Predicting the Effects of Missense Variation on Protein Structure, Function, and Evolution

Predicting the Effects of Missense Variants on Protein Structure, Function, and Evolution

A DISSERTATION PRESENTED
BY
DANIEL MICHAEL JORDAN
TO
THE COMMITTEE ON HIGHER DEGREES IN BIOPHYSICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOPHYSICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2015

Thesis advisor: Professor Shamil R. Sunyaev Daniel Michael Jordan

Predicting the Effects of Missense Variants on Protein Structure, Function, and Evolution

Abstract

Estimating the effects of missense mutations is a problem with many important applications in a variety of fields, including medical genetics, evolutionary theory, population genetics, and protein structure and design. Many popular methods exist to solve this problem, the most widely used of which are PolyPhen-2 and SIFT. These methods, along with most other popular methods, rely on multiple sequence alignments of orthologous protein sequences. Based on the amino acids observed in each column of the alignment, they produce a profile describing how tolerated each amino acid is at each position. They then compare the wild-type and variant amino acids to this profile to produce a prediction.

In practice, these methods are fast, robust, and relatively reliable. However, from a theoretical perspective, they have at least three significant shortcomings:

1. They use effects on selection as a proxy for effects on phenotype and protein structure and function.

2. They treat each position as independent, ruling out most forms of interactions between sites.

3. They do not explicitly model the process of evolution, instead assuming that sequences we observe more or less represent an equilibrium state.

With the recent explosion of sequencing technology, as well as the steady increase of computational power, we are now beginning to have enough data to investigate these simplifications and see how much they really affect the performance of these methods.

In this dissertation, I present three such investigations. First, I describe a modified predictor designed to predict risk for a specific disease, hypertrophic cardiomyopathy (HCM), rather than general seletive effect. This method achieves significantly higher accuracy than methods without such specific domain knowledge. Next, I describe a model of pairwise interactions between sites, demonstrating both statistically and with *in vivo* evidence that approximately 7–12% of disease-causing variants may be mispredicted by these methods due to such interactions. Finally, I describe a hybrid method that uses an alignment-based estimator to inform a parametric model of evolution, resulting in a small but significant improvement in accuracy.

Contents

Dedicated to the memory of my grandparents, Joseph and Florence Jordan and Walter and Helen Klopper. They would have been so proud.

Acknowledgments

It is a bit of a cliché to say that grad school was a difficult journey, but it is undoubtedly true. It has taken its toll — on my bank account, on my mental and physical health, and on my relationships. Still, over the past seven (seven!) years I have done a lot of work of which I am very proud, much of which is contained in this document, and, more importantly, learned and grown a great deal, both scientifically and personally. There are many people I must thank for making that possible.

Thanks to Jim Hogle and Michele Jakoulov, for running the least soul-sucking Ph.D. program I have ever heard of; my advisor, Shamil Sunyaev, for his wide-ranging expertise and faintly hubristic attitude towards science; and Ivan Adzhubey, who spent many hours fixing things I broke and tolerated my attempts to improve on his software.

Thanks to all the professors who have sat on my various committees, namely, Steve Blacklow, Leonid Mirny, Eugene Shakhnovich, Kricket Seidman, Mike Desai, Jun Liu, and Cynthia Morton; to my collaborators at the LMM: Sam Baxter, Matt Lebo, Birgit Funke, and Heidi Rehm; and at Duke: Stephan Frangakis, Erica Davis, and Nico Katsanis.

Thanks to my scientific mentors from the before times, without whom I never would have gotten here: Laurel Harmon, Erik Zuiderweg, Bob Koeppe, Martin Philbert, Doug Stone, and, going all the way back to middle school, Tim Wilson.

Thanks to all the members of the Sunyaev Lab past and present, and especially Chris Cassa and Dan Balick, for many hours of useful conversation and advice and many more of useless conversation and whiskey.

Thanks to Brandy Freitas, my drinking buddy, co-TF, and all around comrade-in-arms in the Biophysics program; and to Julia Liu, Sophie Zaaijer, Amy Xu, and David Ferrero, my crack team of scientist-musicians, who were instrumental (ha ha) to maintaining my sanity.

Thanks to Adam Becker and Vlad Barash, who blazed the trail to the Ph.D. for me and always had a sympathetic ear for grad school frustrations; and to my sometime girlfriend Rachel Carpman, who was a source of great strength for many years, and remains one of my best cheerleaders.

Finally, my deepest thanks and love to my parents Lawrence and Elizabeth Jordan and my brother Jonathan Jordan, from whom I have never felt a lack of love and pride; and my dear, dear friends Max Gladstone, Stephanie Neely, and Marshall Weir, who have loved and supported me unconditionally for the last seven years.

0

Introduction

0.1 Motivation

In recent years, deep sequencing of genomes and exomes has begun to take the place of traditional association studies as the method of choice for probing the landscape of human variation. The growth of next-generation sequencing has brought with it an explosion of rare variants, with every newly sequenced individual carrying hundreds of never-before-seen coding variants[38,141,179].

Confidently assigning functional or phenotypic significance to these variants is a very difficult task. Traditional human genetics methods involving observation in multiple individuals with the same phenotype are not usually feasible for these extremely rare variants. In some genes, as many as half of all known variants are intractible to traditional genetic evidence and considered to have "unknown significance," either due to the variant not having been observed in a large enough sample of individuals or due to the lack of appropriate race-matched controls for these observations[153]. On a small scale, such as for an individual patient, these unknown variants can be addressed with *in vivo* or *in vitro* functional testing. However, it remains unfeasible to apply these experiments to the thousands of variants observed by a large clinical lab or the hundreds of thousands of variants observed in a population-level sequencing study.

To fill this gap, a large and growing collection of *in silico* methods to predict the effects of variants has emerged over the last decade and a half[34,101,144]. These methods have seen a great deal of use in a variety of applications, including prioritization of candidate variants and genes, characterization of fitness effects of variants on a population level, and diagnosis of rare Mendelian disorders[4,7,179]. However, these tools are still generally seen as immature, and are often compared unfavorably to functional analysis and human genetic evidence[71,153,176,205]. One important reason for this is that the accuracy of these methods remains fairly low, and the groups that produce them systematically overestimate the accuracy of their own methods, resulting in a reputation for unreliability[72,183]. Another is that the databases containing variants of "known" effect used to train these predictors are polluted with dubious annotations, which both affects the accuracy of the predictors themselves and contributes to the perception that computational methods are unreliable[27,111].

As whole-exome and whole-genome sequencing become more and more common, the number of variants in need of interpretation is only going to increase, and the need for accurate and trusted tools for variant interpretation is only going to become more acute. Improving the reliability and reputation of these tools and addressing their quirks and faulty assumptions is essential for the fu-

ture of the field.

0.2 The Method

Evolution can be seen as one enormous *in vivo* experiment for evaluating the effects of amino acid changes. Changes are proposed by a stochastic mutation process, and those with negative effects on fitness are rejected by natural selection. Thus the history of evolution contains an enormous amount of information about which changes are tolerated and which are not. We can access part of this history using comparative genomics, by comparing sequences from different species. The oldest and most widely used variant effect prediction methods, SIFT and PolyPhen, have relied on this insight since their introduction in 2001[142,172], and most widely-used methods continue to do so to this day[101].

In general, the fixation or loss of an allele is controlled by the strength of selection against that allele. To be precise, the probability of fixation π_1 of an allele with frequency p is

$$\pi_1(p) = \frac{1 - e^{-2N_e sp}}{1 - e^{-2N_e s}} \tag{1}$$

where s is the strength of selection against the allele* and N_e is the effective population size† [67]. If selection is sufficiently strong ($N_e s \gg 1$), the allele is deterministically lost; alleles with such strong selection acting against them should never be found fixed in any species. On the other end of the spectrum, if selection is sufficiently weak ($N_e s \ll 1$), genetic drift overcomes selection against the

*Note that for the purposes of this discussion, and indeed through the bulk of this dissertation, I am ignoring positive selection. There are tools to detect positive selection, but for the most part positive selection is irrelevant to the problem of variant effect prediction, as alleles under sufficiently strong positive selection fix very rapidly and are unlikely to be found as variants requiring interpretation.

†I won't go into the exact theoretical definition of N_e, but in most cases it can roughly be thought of as the minimum population size at the last population bottleneck. In this case we are not really talking about a single population, but rather a pool of many different populations over the history of evolution. Based on the effective population sizes measured for existing vertebrate species, we can probably assume that N_e for these ancestral vertebrate populations would have been roughly in the range of 10^3–10^6.

allele, and the allele can become fixed stochastically with probability p just as though it were neutral. According to this principle, any allele that is seen to be fixed in any species can be assumed to have only weak selection acting against it.

Even at the most basic level, this principle is remarkably useful for predicting the effects of alleles and the relative importance of sites. For example, suppose we implement the following extremely simple classifier: any allele observed in another species in our multiple sequence alignment (in this case we will use the MultiZ whole-genome orthologous alignment of 100 vertebrate species[103]) is benign, and any allele never observed is pathogenic. Using the HumVar dataset, a dataset of variant annotations based on the SwissVar database and commonly used for training and testing variant effect predictors[22,139], this extremely simple classifier correctly predicts 91% of pathogenic variants and 63% of benign[99]. This should be considered a baseline for many more advanced prediction algorithms; in some sense, the further refinements made by methods like PolyPhen and SIFT only serve to address the 37% of benign variants that are mispredicted by this simple method[‡].

The difference between this method and the basic framework used by methods like SIFT and PolyPhen is that these methods move from the yes-or-no question of "is this amino acid tolerated at this site?" to the concept of a "profile" of amino acid preferences at each site. Moving from a more-preferred amino acid to a less-preferred one is likely to be damaging, while moving in the opposite direction is probably not. The way we typically think about this profile is as the distribution of amino acids likely to be found at the site. It's often referred to as the equilibrium distribution of amino acids, based on the idea that evolution is a stochastic process that we are sampling from when we observe sequences. If the process is at equilibrium, the amino acids we observe should be drawn from this equilibrium distribution. The most naive way of computing this distribution is simply to count up the number of observations of each amino acid — e.g. if there is a position where we ob-

[‡]The 9% of pathogenic variants that are mispredicted by this simple method turn out to be much harder to deal with. They will be discussed in greater detail below; Chapter 2 of this dissertation is aimed at characterizing these variants.

serve 15 sequences with valine, 4 with isoleucine, and 1 with tryptophan, we would naively describe the preference of this amino acid site as 75% valine, 20% isoleucine, and 5% tryptophan.

There are additional adjustments that are commonly made on top of this naive method. One very commonly used adjustment is to add pseudocounts to account for the chemical properties of amino acids and our prior expectations about the variability of sites[80]. For example, in our valine-isoleucine-tryptophan example, we might observe that tryptophan is very chemically different from valine and isoleucine, and its presence therefore suggests that the site is probably somewhat tolerant of a wide variety of different amino acids. Our method might add pseudocounts of other non-observed amino acids to account for this variability. We might also observe that leucine is very chemically similar to valine and isoleucine, and if these two amino acids are both tolerated leucine probably is as well. We could therefore add pseudocounts of leucine to account for the fact that our observations of valine and isoleucine represent implied observations of leucine. Another adjustment is to weight sequences based on their relatedness[171]. This is based on the insight that two very closely related sequences carrying the same amino acid are not really two independent observations of that amino acid, since they are sampled from the same branch of the evolutionary tree.

This profile must then be converted to a prediction. The amino acid profile of a site is a 19-dimensional vector[§], but a prediction of pathogenicity should ideally be a one-dimensional value. Some methods use a simple statistical model to make the conversion; for example, SIFT[142] simply reports the expected frequency of the variant amino acid, while PANTHER-PSEC[180] reports the log likelihood ratio of the wild-type and variant amino acids. Others, such as PolyPhen-2[2] and SNAP[18], use the frequencies from the profile as features in various machine learning classifiers. This approach also allows explicit incorporation of other features, such as sequence context, secondary structure elements, or geometric features extracted from solved 3D structures. In fact, most such methods

[§]There are 20 different amino acid types, but the requirement that the frequencies sum to 1 restricts the 20th value, so there are only 19 independent frequencies.

also incorporate some additional information of this kind. Even with the addition of these features, though, the amino acid tolerance extracted from the multiple sequence alignment remains the most informative predictive feature for all predictors that use it.

Many variations exist on this basic method, using different methods for retrieving sequences and constructing alignments, different models to extract and adjust profiles, different sources of additional annotation, different machine learning methods, and different training datasets of known variants. However, the basic method of using the profile of amino acids observed in evolutionary history to estimate the amino acid tolerance of a given site is still widely used today, 14 years after the release of the original PolyPhen and SIFT methods. Even the latest state-of-the-art methods still use these profile scores as major components of their predictors[25,108]. It may seem strange that there has been no major improvement on this method, considering that it is completely agnostic about biological function and protein structure, which should both be vitally important to determining the strength of selection. Though many methods incorporate features intended to capture these factors, these features rarely add much information on top of the profile score. This information seems already to be encoded in the profile score. After all, in some sense the profile of observed amino acid frequencies is the result of integrating this information throughout the history of evolution.

0.3 Issues

Careful readers of the above may notice a number of places where I have hand-waved away important distinctions or discarded large swaths of established theory. This is not entirely due to my own lack of care. There are a few important issues that most modern methods systematically ignore. Three of these issues will motivate the three subsequent chapters of this dissertation:

0.3.1 Deleterious vs. Pathogenic

One area where the field as a whole lacks clarity is the question of what, exactly, these tools are designed to predict. Common use cases, including most of those mentioned above and most of those proposed by the authors of these methods, primarily focus on predicting *phenotypes*, especially medically relevant phenotypes. However, what these methods really measure is *selection*. In general principle, selection is not necessarily a good proxy for phenotypic effect. There are many severe and medically relevant phenotypes that may not produce very large selective effects, such as those with late onset, reduced penetrance, or pleiotropic effects on selection[198,208]. Conversely, there are many mild or medically irrelevant phenotypes that nevertheless produce large selective effects. For example, a sperm motility defect causing a ~10% reduction in fertility should appear extremely deleterious, despite having minimal noticeable effect on the individual's health.

This conflation obviously has the potential to influence the accuracy of predictions, and it has consistently been remarked on across the entire history of variant effect prediction, from the original paper reporting the PolyPhen method ("only a small fraction of deleterious amino-acid altering SNPs ...lead to total loss of function of the affected protein, and the rest must have relatively mild effects")[172] to the paper reporting the CADD method published last year ("it is at present not possible to precisely calibrate the relationship between ...estimated deleteriousness and the likelihood that a variant is pathogenic")[108]. Nevertheless, most people who use or design these methods do not pay much attention to this problem, since the predictions do, after all, work reasonably well. Undoubtedly they could work better if they made an effort to account for the relationship between selection and pathogenicity, but it's unclear how much better. Additionally, accounting for this relationship probably requires a great deal of domain knowledge about specific genes and phenotypes, and may not be possible on a genome-wide level.

0.3.2 Independence of Sites

It seems reasonable enough to claim that evolution explores the range of allowed amino acids at every site, and that we can reconstruct that range by looking at the sequences output by evolution. However, hidden within this formulation is the tacit assumption that the range of allowed amino acids is defined site-by-site, and that each site has a fundamentally independent profile of amino acid preferences. In fact, the massive *in vivo* experiment of evolution does not just produce legal lists of amino acids, but instead legal protein sequences, each of which folds into a three-dimensional structure and carries out its function as a complete sequence. The two concepts are only equivalent if interactions between sites have minimal biological importance, and the effect of multiple variants together is almost always the sum of their individual effects. There is ample evidence in modern genetics that interactions between sites do exist and are important, though there is some debate over how common they really are in the history of evolution[6,16,36,130].

The specific worry for variant effect prediction methods is that a nonhuman sequence might contain a specific amino acid that would be pathogenic in the context of the human sequence, but is observed in a context that compensates for its pathogenic effect. In this case, a prediction method would be likely to incorrectly predict the variant as benign, because in fact it is observed in nature as a benign variant. This situation, where a variant that causes a disease in human is fixed in another species, is often referred to as a compensated pathogenic deviation (CPD). Several studies have observed and commented on this phenomenon, frequently also observing that roughly 10% of fixed differences between species are pathogenic to one of the species[30,112,115]. However, despite this relatively frequent occurrence, prediction methods generally do not make an effort to account for CPDs. This is generally because there is no way to know *a priori* which interactions are important for compensation. Without specific biological or biochemical knowledge about the protein in question, the relevant compensation could be any position or combination of positions in the entire

genome.

0.3.3 Parametric Methods

The idea that sequences are samples from a stochastic process of evolution is not unique to these prediction methods; in fact, it's a very common way of viewing evolution. There is a large and well-developed set of models that treats evolution as a branching continuous-time Markov process[92]. These models explicitly account for the tree structure of relationships between sequences. They also can in priniciple explicitly account for differences and similarities between amino acids, accepting as parameters both the relative frequencies of different amino acids and the rates of substitution between them. It may then seem odd that variant effect prediction methods do not use these state-of-the-art models, preferring to use the heuristic methods of sequence weighting and pseudocounts[¶]. In fact, there are some methods that do use these models, the most widely-used of which are phyloP[154] and GERP++[41]. They are referred to as "parametric" methods, because they deal with the rate parameters of these models, as opposed to the profile methods, which do not have parameters in the same sense. These methods are often used to identify funcitonal sites and conserved elements, especially in noncoding regions[4,34,79]. However, they are generally not as accurate as profile-based methods for amino acid substitutions[52,108], and therefore are not as widely used in coding sequences.

It is somewhat frustrating that parametric methods don't perform better, since they should in some sense be more correct than profile methods. Especially in cases where reasonably trustworthy phylogenetic trees exist, such as for whole-genome orthologous alignments where we in principle know the relationships of the species, it seems like we are throwing data away by not explicitly modeling these trees. One easy explanation for why these methods don't work as well is that they are estimating the wrong quantity: instead of the profile of tolerated amino acids at a site, they esti-

[¶] These methods are "heuristic" in the sense that they ignore features of the theoretically correct model and use statistical corrections for them instead. However, I do not mean to suggest that the methods are inaccurate or incorrect; in fact, as a general rule, they seem to perform better than the full models.

mate the rate of evolution at the site. This rate parameter should be a good indicator of evolutionary constraint, but it does make some amount of sense that the higher-dimensional amino acid profile might contain more information than a single rate score per site. In principle, it is possible to represent the equilibrium distribution of amino acids in a parametric method — indeed, any such method must deal with this distribution on some level, because it is a feature of the model. The results produced tend not to be too useful, though. Based on my experience experimenting with these models, there appear to be two reasons for this:

1. It is very difficult to deal with the full 20-letter amino acid alphabet in a fully parametric way. It requires a 20×20 substitution rate matrix with 380 independent parameters, far too many to reasonably infer using the 100-sequence orthologous alignments we typically have access to, not to mention the seconds per site or less of computational time we typically have in large-scale prediction tasks.

2. The likelihood surface of the 19-dimensional amino acid frequency vector seems to be very flat. If there are two amino acids that are both clearly tolerated, it is very difficult to say what the "correct" values for their relative preferences are. Is it 50%–50%? 80%–20%? 20%–80%? Profile methods can safely ignore the differences between these scenarios; parametric methods can't, because they have a dramatic influence on parameter values. In a likelihood surface that is high-dimensional and lacks a sharp peak, likelihood methods can easily get stuck in local maxima or fail to collect enough information to overcome the prior.

In the course of my doctoral studies, I have investigated these and other concerns about variant effect prediction methods, with the hope of pointing the way towards improvements in these already well-established methods. The following chapters represent the results of three research projects aimed at addressing the three specific issues described above.

But wait: let us question a holy man,
a prophet, even a man skilled with dreams—
dreams as well can come our way from Zeus—
come, someone to tell us why Apollo rages so,
whether he blames us for a vow we failed, or sacrifice.
If only the god would share the smoky savor of lambs
and full-grown goats, Apollo might be willing, still,
somehow, to save us from this plague.

Homer, *Iliad* book 1[83]

1

Variant effect prediction in the clinic

Clinical professionals have long had the intuition that computational prediction methods do not work well enough for clinical applications[71,153,176,205]. Part of this problem is the actual performance of these methods, which is far below the levels required for clinical use. But a more subtle and pervasive problem is that lack of real clinical validation. Questions like "what is the odds ratio of this test?" simply have no answer, because most of these methods have never been tested

with actual patients in a realistic clinical setting. This is a manifestation of the confusion between pathogenicity and deleteriousness: these methods are reasonably good and consistent at distinguishing neutral and deleterious variants, but their performance on predicting clinical phenotypes is inconsistent, often poor, and difficult to measure in any case.

The study described in this chapter was undertaken in 2009–11 with the aim of addressing this issue. We replaced the usual datasets used for training and validation — lists of putatively benign or pathogenic variants derived from publicly available databases — with a list of variants classified by a clinical genetic diagnostic lab (the Laboratory for Molecular Medicine, LMM) for clinical relevance in a specific clinical phenotype (hypertrophic cardiomyopathy, HCM). This gave us the ability to evaluate the method's actual performance on a specific phenotype, which then let us attempt to improve that performance. The clinical use case is different from the general-purpose predictor in several important ways, which give us the possibility of improving the method's performance:

1. We are attempting to predict a specific phenotype rather than an overall concept of "deleteriousness," which allows us to incorporate specific annotations related to the molecular mechanisms that cause that phenotype.

2. We are focusing on a small number of well-studied genes, rather than trying to make predictions that will work for the entire genome, which makes it feasible to perform manual inspection and curation of alignments and annotations.

3. Our users do not expect instant results for any position in the genome, which allows us to use methods that are more computationally intensive, either by precomputing results for the relevant genes or by allowing the method to run for more time.

4. It is more important that predictions be accurate than that every variant receive a prediction, which allows us to sacrifice coverage to improve accuracy.

Using these insights, we developed a new predictor and measured its accuracy at 92%, a satisfactory level for clinical use. The method we produced remains available online at http://genetics.bwh.harvard.edu/hcm, and has been used as part of LMM's variant assessment pipeline since its completion in 2010. A study on rare variants on the Framingham Heart Study population suggested that our method's accuracy was comparable to that of manual classification by experts, though it did suggest that even manual classification may overpredict causative variants[15]. However, despite our relative success, few published studies since have attempted to create single-phenotype prediction methods. Most groups that are interested in predicting variant effects continue to prefer methods that are useful across a broad range of phenotypes. This is a perfectly sensible preference, since broadly useful methods are easier to evangelize to the medical and scientific communities and may be of greater scientific interest. Nevertheless, our results suggest that there may a hard limit to how accurate these methods can be without accounting for molecular features of specific phenotypes.

The remainder of this chapter originally appeared in *The American Journal of Human Genetics* in 2011[100]. Accordingly it is copyright 2011 by the American Society of Human Genetics. It is reproduced here with permission. My primary contribution to this work was the design and implementation of the new predictive features described in 1.2.3 and the alignment pipeline described in 1.2.4, as well as major contributions to the overall study design. Supplemental materials are available online at http://www.cell.com/ajhg/supplemental/S0002-9297(11)00012-7.

1.1 Background

DNA sequencing is quickly becoming the method of choice for clinical genetic diagnostics. The improvement in clinical sensitivity that sequencing provides over genotyping platforms is invaluable, especially in disorders that show locus and allelic heterogeneity. However, there are also important challenges presented by the use of DNA sequencing, including the difficulty of interpreting novel

sequence variants. There is currently little standardization of variant classification in the genetics community. Most clinics use a combination of traditional genetic methods relying on segregation with the disease in families, frequency in controls, biochemical characterization, and evolutionary conservation at the variant position[159]. This manual classification process is time-consuming and requires significant expert knowledge. More frustratingly, it often fails to produce a classification at all: variants with incomplete or conflicting data are routinely classified as "variants of unknown significance" (VUSs), and no confident classification is reported to the patient or the referring physician. In some genes, these VUSs comprise as many as one quarter to one half of all reported variants[153]. This problem is only getting worse. As next-generating sequencing technologies begin to enter widespread clinical use, the volume of novel variants should be expected to expand by several orders of magnitude. The genetics community must begin to develop robust automated methods to classify novel variants accurately.

There currently exist several computational tools for predicting the functional effects of genetic variants[101,144,184]. However, these tools in general were not designed for clinical use, have not been rigorously tested on individual genes or diseases, and have not undergone any kind of validation against well-curated datasets. Therefore, the sensitivities and specificities of these predictors are in general ill-defined. This lack of proper validation has created the perception among medical professionals that automated predictors cannot be trusted[176]. Consequently, although most geneticists are familiar with these tools, the predictions they produce are typically not formally included in clinical variant classification methods and are therefore not communicated to physicians via clinical reports. Several studies have attempted to address this problem by validating existing predictors against known disease-causing variants, largely arriving at the conclusion that these methods are not yet mature enough for clinical use[53,175,176].

Variant classification pipelines that are considered mature enough for clinical use are generally designed from the ground up with clinical use in mind, and are designed, demonstrated, and vali-

dated using variants classified according to clinical criteria. Examples of such pipelines include the classification procedure currently in use at the the Laboratory for Molecular Medicine (LMM), a clinical diagnostic laboratory in the U.S., and the integrated evaluation of BRCA gene variants that developed from the work of Goldgar et al.[69]. However, fully automated computational predictors are not currently designed in this way. We therefore set out to test whether this methodology could successfully create an automated predictor that would be useful to medical professionals as a tool for classifying novel missense variants. We chose to target one specific disease and a limited number of genes in which disease-causing variants might be found, so that we would be able to generate a high-quality set of manually classified missense variants to use as the gold standard for training and validating our predictions. We also hoped that focusing on a limited number of functionally related genes would allow us to identify common features of these genes and common mechanisms of disease in these genes, which would help us to make our predictor more accurate.

The disease we chose was hypertrophic cardiomyopathy (HCM), an autosomal dominant disease of the myocardium (heart muscle) with an incidence of roughly one in 500 individuals and a largely genetic basis[190]. Variants in over 20 genes are associated with HCM, with over 900 unique variants reported in the literature, and sequencing of many of these genes can be ordered for clinical testing in CLIA-approved laboratories. The vast majority of pathogenic variants are found in eight genes that encode for units of the cardiac sarcomere, a contractile protein complex in the heart: beta-cardiac myosin heavy chain (*MYH7*), cardiac actin (*ACTC1*), cardiac troponin T (*TNNT2*), alpha-tropomyosin (*TPM1*), cardiac troponin I (*TNNI3*), cardiac myosin-binding protein C (*MYBPC3*), and the myosin light chains (*MYL2*, *MYL3*). Sequencing of these genes yields a high number of novel variants, mainly due to the high prevalence of private familial variants. Roughly 50% of probands tested have a disease-causing variant in one of these genes and approximately 80% of those are in *MYH7* and *MYBPC3*[158] LMM unpublished data. Missense variants represent nearly all such variants detected in *MYH7* and 35% of those in *MYBPC3*. Missense variants exerting dominant neg-

ative effects on the sarcomere structure represent the vast majority of all variants. The notable exception is *MYBPC3*, where missense variants constitute only about 35% of all variants, the remainder being splice, nonsense or frameshift variants leading to loss of function. At the time of this study, the Laboratory for Molecular Medicine had identified over 700 variants in HCM-related genes over five years of testing, over half of which were novel at the time of reporting and over half of which were missense changes. We performed a systematic manual classification of these variants, producing a final dataset of 74 missense variants with extremely confident manual classifications. Using these 74 variants as our gold standard, we then set out to develop and validate a novel computational method that could predict the pathogenicity of any variant in these six genes.

1.2 Methods

We created a computational method to predict the pathogenicity of a novel variant in any of the six genes we chose to screen for HCM mutations. Our method, like other existing methods[17,142,143,203,204] and particularly the recently developed algorithm PolyPhen-2[2], integrates phylogenetic and structural information from several heterogeneous sources using a probabilistic classifier. However, unlike these methods, it exploits the narrow focus on six specific genes known to contain variants that cause the disease to improve the prediction strategy significantly. Also unlike these methods, it uses variants classified according to clinical criteria of pathogenicity to train the probabilistic classifier. The selection and classification of these variants, the features used for classification, and the training and validation of the classifier are all described below.

1.2.1 Selection of target genes

HCM is caused primarily by variants in eight genes encoding protein subunits of the cardiac sarcomere. We initially attempted to use all eight genes to develop our predictor. However, after con-

structing our dataset (see 1.2.2 below), we examined the distribution of variants and found that the final dataset contained no variants in *ACTC1* and only one in *MYL3*. We discarded these two genes and built our classifier around the remaining six (*MYH7*, *TNNT2*, *TPM1*, *TNNI3*, *MYPC3*, and *MYL2*).

1.2.2 Manual classification of HCM variants

We relied on LMM's standard variant assessment pipeline to create our dataset of manually classified variants. To ensure unbiased training and testing of our computational method, we excluded from manual classification information that was accessible to the method such as evolutionary conservation or structural data, even though this information is currently used in the pipeline. Each variant recieved a classification of "Pathogenic," "Likely Pathogenic," "Benign," "Likely Benign," or "Unknown Significance" (VUS). The basic decision process we used is described below and shown in Figure 1.1.

Pathogenic. Variants with a minimum of five informative meioses supporting familial co-segregation with HCM, absent in healthy controls, and/or having strong functional data are classified as pathogenic. In HCM, informative meioses typically only include individuals who are positive for both phenotype and genotype. This level of stringency is required due to the highly variable expressivity and reduced penetrance, which makes individuals without the phenotype largely uninformative, regardless of their genotype.

Likely Pathogenic. The minimum requirement to classify a variant as likely pathogenic is absence from race-matched controls or a large cohort of race-matched probands. The LMM has previously sequenced sarcomere genes in over 1000 HCM probands of European ancestry. Absence from this cohort was accepted in lieu of healthy control data because it serves to set a maximum population frequency of one per the total number of probands tested. Novel

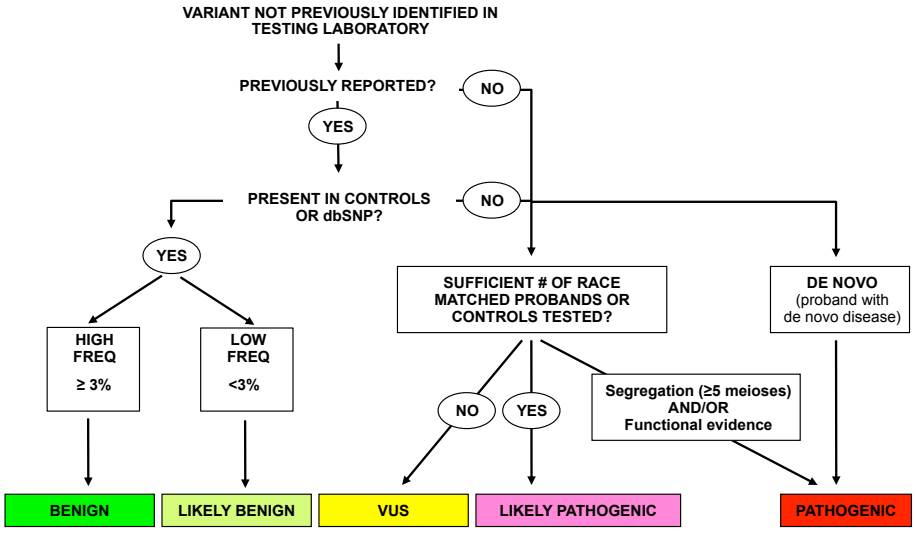


Figure 1.1: Process used to classify variants at the LMM. This process is described in detail in 1.2. We treat the "Pathogenic", "Benign", and "Likely Benign" categories as high-confidence classifications for the purposes of training the automatic classifier.

variants detected in minority populations are therefore often classified as of "Unknown Significance" due to the lack of control cohorts or large proband data sets.

Benign or *Likely Benign.* Variants that are frequent in the general population (at least 3%) are classified as "Benign." Variants present in controls at frequencies below 3% and without other suspicion for pathogenicity are classified as "Likely Benign."

Unknown Significance (VUS). This class commonly includes variants for which there is insufficient evidence to classify the variant in any of the other four categories, or variants for which the evidence is conflicting.

Figure 1.2 shows the distribution of variants by the classification category in our database.

After applying these criteria to the complete set of variants collected by LMM, we filtered the resulting dataset to exclude unconfident predictions. We excluded variants in the "Likely Pathogenic" category, considering the classification for this category not to be stringent enough. We also excluded variants in the "Unknown Significance" category, since this category carries no clinical or biological significance. This left us with 41 "Pathogenic" variants, which we treated as truly pathogenic, and 7 "Benign" and 26 "Likely Benign" variants, all of which we treated as truly benign. These 74 variants became our gold standard for validation of our predictor. The complete list of 74 variants is shown in Supplemental Table S1.

There is a possibility that the manual method of variant classification may have selected variants resulting in the most severe phenotypes, such as those seen in early onset cases, which may reduce the utility of our classifier for less severe variants. To investigate this possibility, we used the age at which an individual was tested as a proxy for age at onset. The distribution of ages of all probands tested is roughly trimodal with clear peaks at less than 1 and 15 years of age and a broad distribution centered around 50 years of age (Supplemental Figure S1). The distribution of pathogenic variants in this population follows a similar distribution with pathogenic variants detected across a wide range

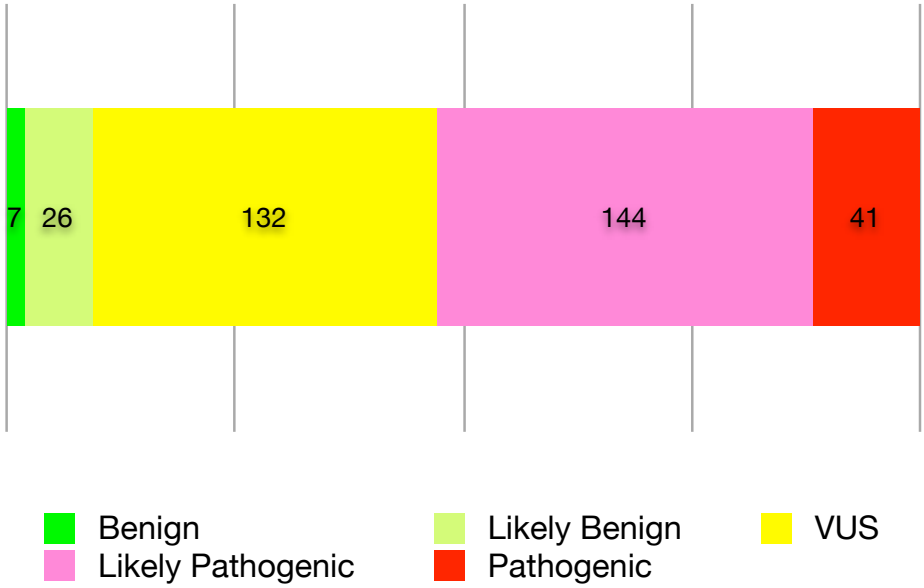


Figure 1.2: Distribution of variant pathogenicity. We categorized 350 missense variants in six genes according to the criteria described in Figure 1.1. The three categories "Pathogenic," "Benign," and "Likely Benign" were treated as high-confidence classifications and used as training data for our classifier (enumerated in Supplemental Table S1).

of age groups tested. If we were indeed selecting for only the most severe, early onset phenotypes, we would expect pathogenic variants to be over-represented in newborns and teenagers and to be absent in late-onset cases. This does not appear to be the case and we are confident that our training set does not only consist of pathogenic variants that lead to high penetrance, early onset disease.

1.2.3 Predictive features

We used four features in the final predictor. These features are described below.

PolyPhen-2 prediction. Our first feature was a prediction made by the existing method PolyPhen-2[2]. PolyPhen-2's predictions integrate several sources of phylogenetic and structural information using a Naive Bayes classifier. Its output represents a general-purpose prediction, made without knowledge of the specific disease under consideration. The PolyPhen-2 software reports a score ranging from 0 (neutral) to 1 (damaging), which represents the confidence of its internal classifier. We used this integrated score as a single feature in our predictor.

MrBayes substitution rate score. Our second feature was the rate of evolution for each site in each gene. We computed this using the Markov Chain Monte Carlo (MCMC) algorithm in the MrBayes software package[161]. This score took several days of computer time to calculate for all six genes, and would not have been feasible to calculate for a genome-wide dataset.

Examples of the MrBayes instruction files we used are available as Supplemental Figure S2. We used a function that infers site-specific evolution rates and includes them in the program's output. MrBayes reports the rate at positions with insufficient alignment depth as 1.000, so all scores of exactly 1.000 were treated as missing data. We normalized this rate so that the mean rate for each gene was 1.000.

Coiled-coil score. Our next two features took advantage of specific properties of the six target genes. Four of the six target proteins had significant coiled-coil regions: *MYH7*, *TNNI3*, *TNNT2*,

and *TPM1*. We used the COILS2 software to predict the tendencies of the wild-type and mutant sequences to form coiled coils[125,126]. Variants that significantly change the coiled-coil tendency of the sequence are likely to interfere with protein function.

For each of the four proteins, we downloaded annotations from SMART to determine the locations of coiled-coil regions[120]. For any variant in a coiled-coil region, we ran COILS2 on both the wild-type and variant sequences of the coiled-coil region that contained the variant. COILS2 outputs a score indicating coiled-coil tendency for each residue in the input sequence, with each score depending on the entire sequence. The feature we used in the final predictor was the magnitude of the largest single-residue change.

Protein structure comparison score. Four of the six target proteins are contractile proteins studied in multiple conformations (*MYH7* and *MYL2* in ATP, ADP and nucleotide-free states; *TNNI3* and *TNNT2* in Ca^{2+} activated and Ca^{2+} free states). For these four proteins, we measured the motion of each residue between the two conformations. Highly mobile residues were considered functionally important to the conformational change, while highly immobile residues were considered structurally important. Intermediately mobile residues were scored as unimportant. We measured the size of each residue's motion by comparing the displacement of the residue to the expected probability distribution of displacements under random thermal motion.

We used two sets of structures to compute this score. One was a set of six structures of a three-chain scallop myosin complex, consisting of the myosin heavy chain (corresponding to *MYH7* in human heart muscle) and the two myosin light chains (corresponding to *MYL2* and *MYL3* in human heart muscle)[82,87]. One of these structures was not bound to a nucleotide (PDB ID 1KK7), two were bound to ADP analogs (PDB ID 1KK8 and 1B7T), and three were bound to ATP analogs (PDB ID 1KQM, 1KWO, and 1L2O). The other set of

structures was a pair of structures of a three-chain chicken troponin complex, consisting of troponin I (corresponding to *TNNI3* in human heart muscle), troponin T (corresponding to *TNNT2* in human heart muscle), and troponin C (corresponding to *TNNC1* in human heart muscle)[188]. One of these structures was activated by calcium ions (PDB ID 1YTZ), and the other had no calcium bound to it (PDB ID 1YV0).

We performed pairwise comparisons between structures that represented the same molecule in different biological states. Pairs of structures that represented the same biological state (such as 1KK8 and 1B7T, which both represent the ADP-bound state of myosin) were excluded, under the assumption that differences between these structures would represent differences in the experimental preparation rather than a meaningful conformational change. We aligned each pair of structures with LovoAlign and measured the displacement between the alpha-carbons of corresponding residues[129].

The variance in the position of an atom in a crystal structure is given by

$$\sigma^2 = \frac{B}{8\pi^2} \tag{1.1}$$

where B is the crystallographic temperature factor for the atom. We computed this variance for the alpha carbon of each residue, estimating B as the average of the reported temperature factor for that atom across the two crystal structures. We used Student's t-test to compare the squared displacement of the atom with its expected variance. This produced a p-value for the observed squared displacement, with numbers close to 0 representing motion much smaller than expected, numbers close to 1 representing motion much larger than expected, and numbers close to 0.5 representing the expected amount of motion. Finally, scores below 0.5 were subtracted from 1, so that a higher score would consistently represent a more important residue.

The human genes were aligned to the structures using BLAST. Each residue in the human sequence was scored the same as the residue it aligned to. Residues that did not align to the structures were not given a score. Only 84 human residues failed to align to the structures, which represents 3.2% of all positions in the four proteins to which we applied this score.

1.2.4 Multiple sequence alignments

Both PolyPhen-2 and the MrBayes score described above use comparative sequence analysis as a source of phylogenetic information. These methods take as input aligned sequences of multiple homologous proteins, and their predictive values critically depend on the quality of the multiple sequence alignments used. Existing computational methods, including PolyPhen-2 and SIFT, rely on automated pipelines to construct multiple sequence alignments[2,142,143]. We used the standard automated alignment pipeline provided by PolyPhen-2, but since we only needed to construct six alignments, we were able to inspect and adjust each alignment manually.

We noticed in our manual inspection that some of the automated alignments were of very poor quality. The worst alignments were for the two proteins that were most highly represented in our data set, *MYBPC3* and *MYH7*. These proteins have numerous homologs at the domain level, arising from the multiple immunoglobulin domains of *MYBPC3* and the highly conserved myosin motor domain of *MYH7*, and the multiple sequence alignments produced using automatically selected homologs are therefore of poor quality. We created new alignments for *MYBPC3* and *MYH7* by manually removing problematic sequences from the automatically generated alignments. This approach allowed us to tune the alignments manually while still taking advantage of PolyPhen-2's automatic filtering of poor alignments and incorrect sequences. The alignments were very deep to begin with, allowing us to remove a large number of sequences without the alignments becoming too shallow to use.

The sequences we removed from the alignments were those that appeared to have only domain-

level homology to the target sequences and/or did not appear to have a sufficiently similar function to the target sequences. In other words, we attempted to create an alignment for *MYBPC3* that consisted only of forms of myosin binding protein C from various tissues and organisms, and an alignment for *MYH7* that consisted only of forms of myosin heavy chain from various tissues and organisms. The resulting alignments were used as input to the PolyPhen-2 classifier and to MrBayes. The sequences used are listed in Supplemental Tables S3–S6, and the resulting alignments are shown in Supplemental Figure S3.

1.2.5 Training and validation

We trained the classifier on the manually-curated set of 74 missense variants in 6 genes. For each variant in the training set, we computed the four features described above (PolyPhen-2 prediction, MrBayes substitution rate score, coiled-coil score, and protein structure comparison score). The values of each feature for each variant can be found in Supplemental Table S2. The training algorithm (Supplemental Figure S4) aims to maximize accuracy of classification while keeping the required level of coverage. To avoid overfitting, the training algorithm uses ten-fold cross-validation (Supplemental Figure S5). This method splits the training data into 10 parts (6 parts of 7 samples, 4 parts of 8 samples), trains the classifier on 9 training parts and tests it on the remaining 1 testing part. It then repeats the split-train-test procedure 10 times, each time with a different part of the data used for testing. In order to account for the different results that would be produced by using different random divisions of the data in this process, we ran 1,000 iterations of ten-fold cross validation, using a different random division of the data each time. We also tested the final classifier using a leave-one-out cross-validation strategy. The classifier assigns a prediction of "Pathogenic," "Benign," or "No Call" to each variant. The "No Call" prediction is given to variants the classifier cannot predict confidently. This category is included so that we can improve the accuracy (fraction of variants predicted correctly) of our confident predictions by sacrificing coverage (fraction of variants predicted

as either "Pathogenic" or "Benign")[153].

1.2.6 Feature selection

To verify that each of these four features made an important contribution, we constructed four incomplete classifiers, each one missing one of the four features. We performed validation on each of these classifiers as described above, and performed a random permutation test to show that the complete classifier had higher accuracy than each of the incomplete classifiers. We performed 10^6 permutations, so that the minimum P-value we could find was 10^{-6}. Out of our four features, only the PolyPhen-2 score had a one-sided P-value greater than this minimum, with $p = 0.0544$; the other three features all had one-sided P-values less than 10^{-6}. We also performed the same test to establish that using manual alignments instead of automatic alignments improved the score, and found that it did with one-sided P-value less than 10^{-6}. Figure 1.3 shows the distributions of accuracies for each set of features in 1,000 runs of cross validation.

In addition to the four features in our final classifier, we also tried replacing PolyPhen-2 with the similar tools SIFT and PANTHER[142,143,180,181]. We found that each performed comparably to PolyPhen-2, though the classifier with PolyPhen-2 performed very slightly better than either, again with one-sided P-values less than 10^{-6}. Interestingly, though PolyPhen-2, SIFT, and PANTHER were each far more informative individually than any other single feature, each made by far the least individual contribution to the full four-feature classifier that included it. Evidently, the other three features together contain enough information to make the PolyPhen-2, SIFT, or PANTHER score largely redundant.

We also investigated the effect each feature had on coverage. This was of particular concern for the structure pair score and the coiled-coil score, each of which is missing entirely from several genes and regions, which could reduce the predictor's ability to make confident classifications in these regions. We found that both the structure pair score and the coiled-coil score actually increase the cov-

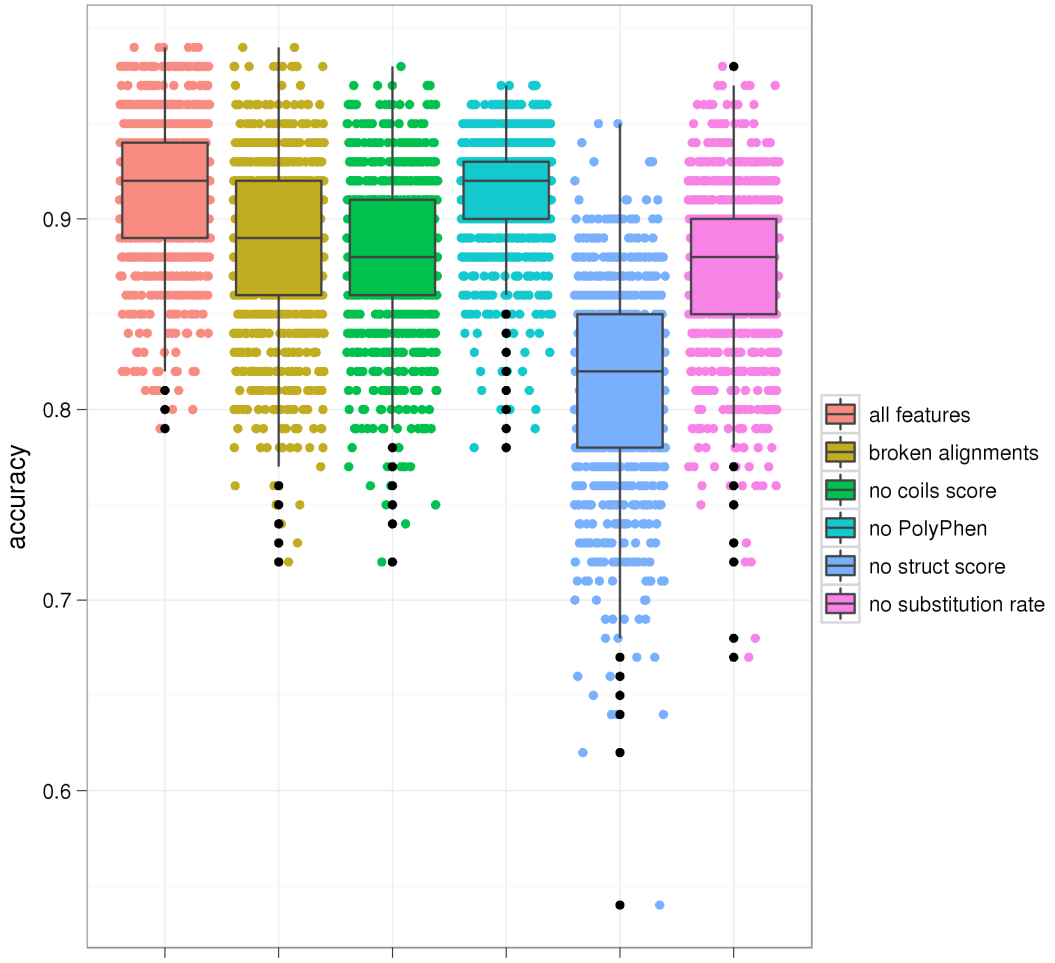


Figure 1.3: Feature selection experiment. Each column shows the distribution of accuracies in 1,000 runs of cross-validation for a classifier built with a different set of features: "all features" represents the final four-feature classifier with manual alignments, "broken alignments" represents the four-feature without automatic alignments, and each of the other four columns represents a three-feature classifier missing the specified feature. Box plots show lower and upper quartiles (50% confidence intervals), and whiskers show 1.5 IQR ranges. The addition of each feature appears to improve the classifier, which is confirmed by a Mann-Whitney test.

erage, while neither of the other features has a significant effect. This suggests that it is rare for a variant that could be scored confidently with the PolyPhen and substitution rate scores to be demoted to "No Call" because it is missing one or both of the other features. In other words, the coiled coil and structure pair scores tend to increase confidence where they are present rather than decreasing it where they are absent.

1.3 Results

1.3.1 The prediction method

We created an automated method to predict the pathogenicity of missense variants in six genes known to contain variants that cause HCM. In designing this predictor, we set out to take advantage of the fact that we were focusing on a small set of functionally related genes to improve our predictions. We identified two ways to accomplish this: first, by exploiting unique structural and biochemical properties of the six target genes, and second, by applying more rigorous methods that would be difficult to implement for large numbers of genes. With these principles in mind, we developed a total of three predictive features, which we used in conjunction with the existing PolyPhen-2 classifier[2]. Two of these features reflect specific structural properties of sarcomeric proteins. One scores the effect of amino acid change on coiled-coil regions, while the other scores the importance of the mutated residue to functionally important conformational transitions in ATP and Ca^{2+} binding domains. The remaining feature is an estimated rate of evolution at the variant position. This feature was extremely time-consuming to compute and would not have been feasible to apply to a genome-wide dataset. It also was computed from manually adjusted multiple sequence alignments of homologous sequences, which required human intervention to produce. These same manually adjusted alignments were also used as input to PolyPhen-2, improving its performance. We combined these three features and the PolyPhen-2 score using support vector regression, with our set of 74 manually

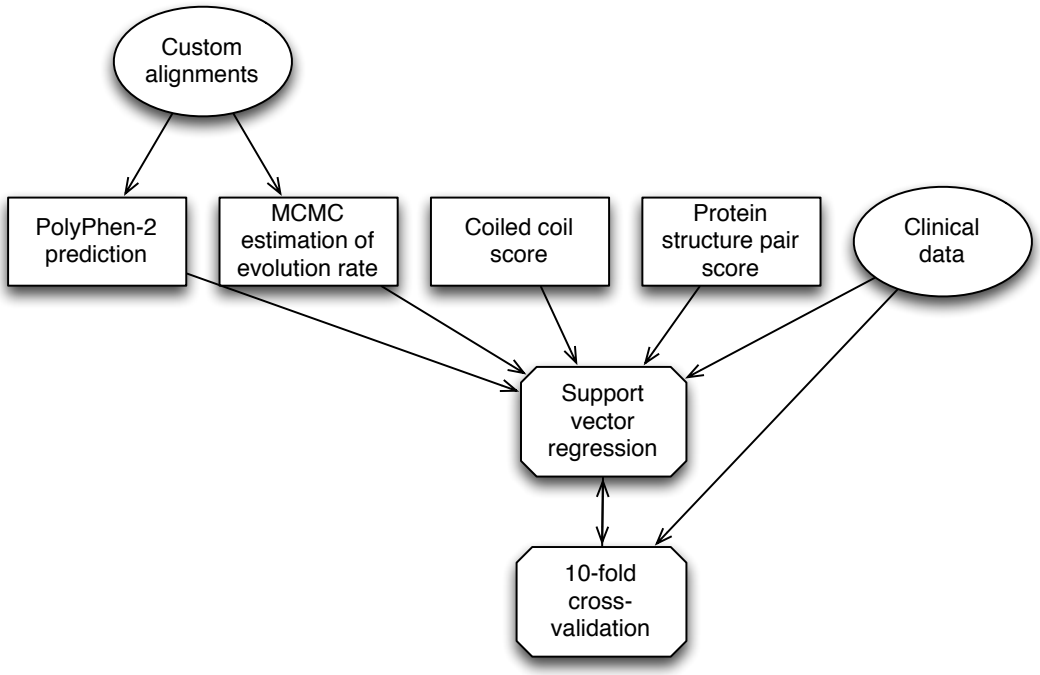


Figure 1.4: The automated prediction process. For each variant, we computed four features and combined them using support vector regression. We trained this classifier on the high-confidence variants classified with clinical data, and validated the classifier against the same data using ten-fold cross-validation.)

classified variants as a training set. The complete method is presented graphically in Figure 1.4.

We also experimented with a small number of alternative features. The most notable among these were a different estimate of the rate of evolution computed using a genomic alignment of 46 vertebrate species, and several of the individual phylogenetic scores used as predictive features in PolyPhen-2. Addition of these features did not improve the performance of the predictor.

1.3.2 Validation of the method against manually classified variants

Given the small size of our gold standard dataset (74 variants), the choice of training and validation method was important. Because we had so few variants, it was not feasible for us to use the

simplest validation method of splitting the dataset in half and using one half for training and the other for testing. Instead, we applied ten-fold cross validation, which is the accepted procedure in such cases (see 1.2.5). We ran this validation process a total of 1,000 times to obtain median results and confidence intervals. Figure 1.5 shows the results of this validation for six different classifiers at different levels of coverage and accuracy. We used the bottom row, highlighted in red, as our final classifier. The method predicts each variant as "Pathogenic," "Benign," or "No Call," with the "No Call" result meaning that the predictor is not sufficiently confident to report a prediction. The median accuracy for covered variants for the most accurate classifier (the fraction of correct predictions out of all "Pathogenic" and "Benign" predictions, while disregarding "No Call" results) was 92%, with a 95% confidence interval of 83%–98%. The median coverage (the fraction of variants that were predicted as either "Pathogenic" or "Benign"), was 57%, with a 95% confidence interval of 49%–64%; in other words, the median classifier reported "No Call" for 43% of variants. The median sensitivity for covered variants (the fraction of variants manually classified as pathogenic that were predicted as "Pathogenic," excluding those predicted as "No Call") was 94%, with a 95% confidence interval of 83%–98%. The median specificity for covered variants (estimated as the fraction of variants manually classified as benign that were predicted as "Benign," excluding those predicted as "No Call") was 89%, with a 95% confidence interval of 83%–98%. The median odds ratio for a prediction of "Pathogenic" (the odds of a pathogenic variant being classified as "Pathogenic" divided by the odds of a benign variant being classified as "Pathogenic") was 10, with a 95% confidence interval of 4.0–infinity (no upper bound could be set since more than 5% of trials had no false positives). The median odds ratio for a prediction of "Benign" (the odds of a benign variant being classified as "Benign" divided by the odds of a pathogenic variant being classified as benign) was 9.9, with a 95% confidence interval of 4.6–21. Leave-one-out cross validation also resulted in highly similar estimates of all these quantities.

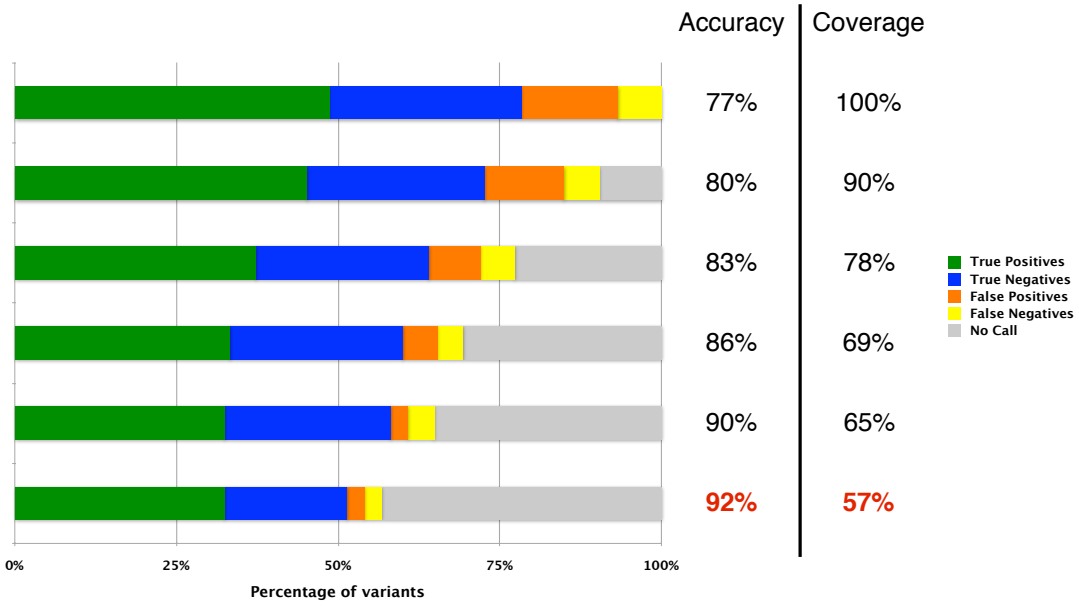


Figure 1.5: Results of cross validation. Rows contain median ten-fold cross validation results for the gold standard dataset at different levels of coverage. Horizontal bars correspond to different levels of coverage and median validation coverage and accuracy levels are indicated. 'True Positives' are variants manually classified as "Pathogenic" that our method predicted as "Pathogenic." 'True Negatives' are variants manually classified as "Benign" or "Likely Benign" that our method predicted as "Benign." 'False Positives' are variants manually classified as "Benign" or "Likely Benign" that our method predicted as "Pathogenic." 'False Negatives' are variants that manually classified as "Pathogenic" that our method predicted as "Benign." 'Uncovered' are variants without a prediction ("No Call"). The bottom-most coverage level, indicated in red, was used for our final predictor.

1.3.3 Comparison with general-purpose methods

Since our predictor bases its predictions in part on predictions of the existing general-purpose method PolyPhen-2, we investigated whether our predictor was a significant improvement over the PolyPhen-2 predictor without our modifications and other general-purpose methods. In order to investigate this, we tested PolyPhen-2, SIFT, and PANTHER on the same dataset. We applied the same ten-fold cross-validation method with each of these three scores as the only predictive feature. We found that all three general-purpose scores had comparable performance on this dataset: PolyPhen-2's median cross-validation accuracy was 70% (95% confidence interval 60%–77%), SIFT's was 74% (95% confidence interval 64%–83%), and PANTHER's was 68% (95% confidence interval 56%–79%). All of these estimates are much lower than the accuracies reported for these methods, which may reflect features of this dataset. Our specialized predictor, on the other hand, had a median accuracy of 92% (95% confidence interval 83%–98%), as reported above. A permutation test showed that all three general-purpose predictors performed worse than our specialized predictor, with one-sided P-values of less than 10^{-6}.

1.3.4 Predictions for variants without confident classifications

The ultimate goal of our predictor is to provide accurate predictions for variants that are not confidently classified by manual methods. This will not be possible if there is some systematic biological difference between the confident and unconfident classifications, such as a difference in penetrance, severity, or mechanism of disease. To determine whether this is the case, we applied our method to a low-confidence dataset, the set of missense variants that did not meet the confidence criteria to be manually classified as truly pathogenic or benign (Figure 1.6). Of the missense variants manually classified as "Likely Pathogenic," 80% of those for which a prediction was made were predicted as "Pathogenic." This is consistent with the expectation that most of these "Likely Pathogenic" vari-

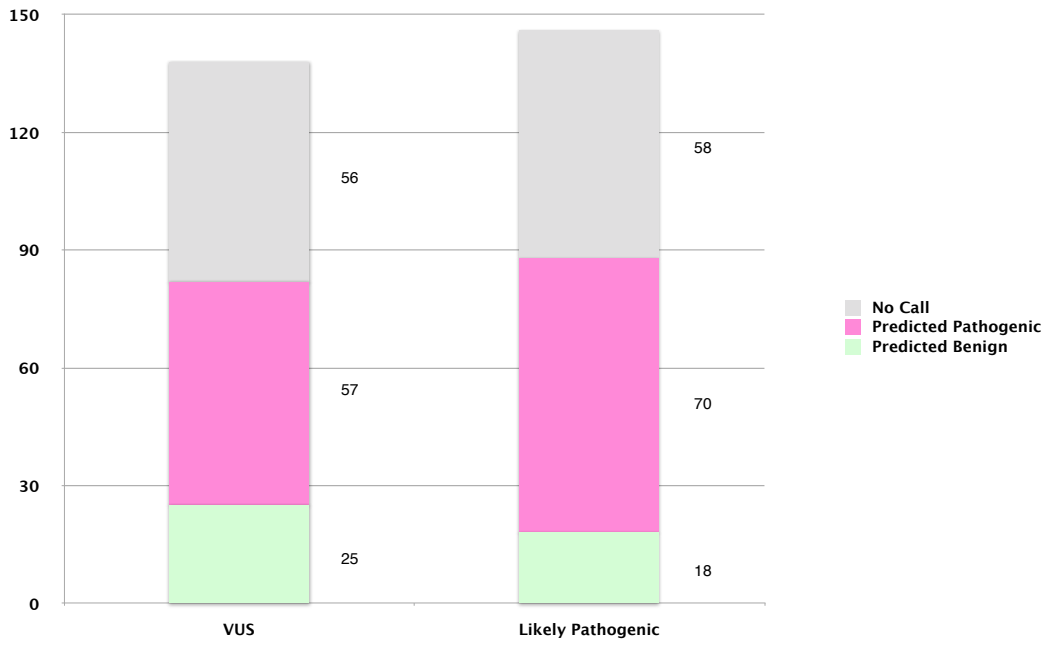


Figure 1.6: Results for low-confidence dataset. Columns indicate, for each class of variants, the number of predictions in predicted categories produced by the final classifier.)

ants are indeed pathogenic. It is also consistent with the expectation that the fraction of variants predicted as "Pathogenic" in this set is lower than for variants manually classified as confidently pathogenic. Among variants manually classified as "Unknown Significance," 70% of those for which a prediction was made were predicted as "Pathogenic." Since these variants have been identified in individuals diagnosed with HCM, there is a higher *a priori* likelihood that they are indeed pathogenic, although we have no way of knowing what the true fraction should be. The fraction of variants predicted as pathogenic remains lower in the "Unknown Significance" set than in the "Likely Pathogenic" set, which is consistent with what would be expected.

We also used the low-confidence dataset to generate an independent estimate of the predictor's coverage. We found that the predictor made a prediction for 60% of low-confidence variants, which

is well within the confidence interval of 49%–64% for the estimated coverage on the gold standard variants.

Discussion

We developed and clinically validated an automated method to predict the pathogenic effect of missense variants that might cause HCM. Unlike current commonly used methods, our predictor has been validated against high-confidence manually curated data. This enabled us to estimate its specificity and sensitivity for the specific task of predicting HCM mutations, which will allow its predictions to be incorporated into clinical reports to health care professionals as one piece of evidence supporting a variant classification. Although this tool adds little for variants whose clinical significance is already supported by strong genetic and/or functional data, it will add value for those variants that had little or no prospect of ever being supported by solid family studies or large scale healthy control studies. Importantly, our classifier is particularly helpful for variants identified in minority populations, where healthy control cohorts, one of the pillars of traditional variant classification, are typically unavailable.

To maintain high accuracy, it was necessary to sacrifice coverage, i.e. the proportion of variants for which a prediction is made[153]. As shown in Figure 1.5, an increase in coverage is accompanied by a rapid decline in accuracy. A method attempting to predict every variant as either pathogenic or benign could not achieve levels of accuracy acceptable for clinical use. We estimated the coverage of our predictor at 57%, with a 95% confidence interval of 49%–64%. We believe this level of coverage is still above the threshold of clinical usefulness. For comparison, note that out of 350 LMM missense variants in the six target genes, only 74 met the criteria for high-confidence manual classification, giving the manual classification process a coverage of only 21%. Note also that our method covers a different set of variants than the manual classification process, including 59% of the variants that the

manual classification classifies as "Unknown Significance."

The most important limitation of our automated prediction method stems from the size of the training data set. In general, training on small data sets may lead to overfitting of automated classifiers. An overfit classifier may be highly accurate on the training data but much less accurate on new data. We applied several safeguards against overfitting during training and validation. These included limiting the number of features in the classifier, using only features that we expected *a priori* to be informative, and performing cross-validation to calibrate parameters and estimate accuracy. In this way we hope we have avoided excessive overfitting in our final predictor.

It is important to point out that this method may not accurately predict the effect of those missense variants that exert their effect partially or fully though affecting mRNA splicing. This is true for all currently available tools of this kind, including PolyPhen-2, SIFT, and others. For example, the MYBPC3 Glu258Lys variant was confidently manually classified as "Pathogenic" but was incorrectly classified as "Benign" in several runs of cross validation (though not in the final predictor). Many MYBPC3 variants affect splicing and there is evidence that the Glu258Lys variant is disease-causing via this mechanism. The underlying cDNA alteration is c.772G>A, which affects the last base of exon 6. This position is known to be part of the splice consensus and 5 different splice predictors (SpliceSiteFinder-like, MatEntScan, NNSPLICE, GeneSplicer and Human Splice Finder; see Supplemental Figure S6) predict an impact on splicing. This is supported by evidence showing that this may result in skipping of exon six[5,128]. Therefore, the conservation of the nucleotide and not the amino acid at this position is essential, possibly explaining a misprediction by our predictor. This is a limitation of this method and clearly lends itself to future improvement and generation of tools that incorporate a splice assessment.

It is also important to point out that clinical laboratories are typically aware of this limitation. Novel variant assessment is a lengthy and complex process that relies on a large collection of different computer tools in combination with traditional genetic evidence such as familial segregation

with disease and absence from race-matched healthy controls. All evidence is taken into account to synthesize a final probability for pathogenicity. In our laboratory, a splice assessment is performed for every variant, regardless of whether it changes an amino acid or not, and a benign prediction by this predictor would not lead to a final classification as "Benign", particularly not for genes where pathogenic splice variants are known to be common.

This example that illustrates that this predictor or any other predictor developed with this methodology should not be used as a sole foundation for a diagnosis but rather be used in combination with other lines of evidence in agreement with ACMG and IARC recommendations[153,159]. We envision future development of a single probabilistic classifier that would automatically combine heterogeneous factors such as familial segregation, frequency in controls, functional evidence and computational predictions following early work in this area.

1.4 Conclusion

We have addressed the problems that prevent automated predictors from being widely used in genomic medicine by developing a custom-tailored predictor specifically designed for clinical use. Our analysis suggests several important considerations that can increase the accuracy of computational methods. Manual adjustment of multiple-sequence alignments and time-consuming computational methods of molecular evolution are feasible when focusing on a small set of genes and may improve predictions that use comparative sequence analysis. Exploitation of specific structural properties of proteins also becomes feasible when focusing on a specific disease. Most importantly, a highly accurate manually-curated dataset is necessary to train and validate an accurate predictor, and this level of validation enables clinical laboratories to include it as part of their variant assessment processes. Where previous studies have concluded that existing tools are not mature enough for clinical use, we believe that our tool is ready for clinical use now, in combination with other sources of information.

Our collaborating clinical laboratory, the Laboratory for Molecular Medicine, has already begun to use our predictor as a source of information about HCM variants, and we look forward to helping additional laboratories do the same. Our study focused on HCM, but we believe that our approach is general and that analogous methods can be constructed for many other diseases where genetic testing is an important part of the diagnosis. In the future we expect to work with additional laboratories and on additional diseases to expand the use of automated predictors in genomic medicine and simplify the problem of interpreting novel variants.

Who are you? where are you from? your city? your parents?
I'm wonderstruck—you drank my drugs, you're not bewitched!
Never has any other man withstood my potion, never,
once it's past his lips and he has drunk it down.
You have a mind in you no magic can enchant!

Homer, *Odyssey* book 10[84]

2

Compensation of disease alleles

It is well established that human disease mutations can become fixed in other species due to compensations present elsewhere in that species' genome. The fixation of pathogenic mutations due to compensation is referred to as compensated pathogenic deviation (CPD), a name coined in the 2002 paper by Alexey Kondrashov et al. that first described the phenomenon[112]. This paper established that approximately 10% of fixed differences between two species are pathogenic

to one of them, a result later reproduced in flies[115]. Many other studies since have recognized the phenomenon and examined the properties of these compensations[8,30,55,60].

Despite broad recognition that pathogenic variants can and do appear in other species, existing methods to predict the effects of variants mostly ignore this phenomenon, assuming that a variant observed in another species is likely to be benign[101]. There are good reasons for this approach, of course: with no biochemical or biological insight into the specific genes and phenotypes involved, there is no obvious way to go about detecting compensation, especially considering that we expect about 90% of variants appearing in other species to be truly benign. Still, it is a little troubling that no previously published study has attempted to estimate the impact CPDs have on the accuracy of prediction methods[*] or to propose a framework for detecting CPDs.

Using the recently available whole-genome orthologous multiple sequence alignment of 100 vertebrate sequences from UCSC[103] and the ever-increasing number of variants with annotated pathogenic effects, we created a dataset of likely CPDs that can be used to address both of these questions. To the first point, we estimated the prevalence of CPDs among known human disease mutations in the range of 7%–12%. To the second, we observed a different distribution among species for neutral variants and CPDs, leading us to develop a statistical model of the fixation and genetic structure of CPDs. We used this model to propose an experiment design for detecting CPDs *in vivo*, and produced evidence for three specific cases in a zebrafish model. We also developed an *in silico* predictor based on our model. As mentioned above, it remains true that any variant seen in another species is far more likely than not to be benign; however, our predictor can at least distinguish between variants that are *likely* to be benign and variants that are *almost certain* to be benign, effectively adding additional levels of gradation to the "benign" classification. This classifier is available

[*]Note that the 10% number from Kondrashov et al. is the number of fixed differences between species that are pathogenic, while the impact on prediction accuracy is determined by the reverse of this number — the number of pathogenic variants that appear in other species. Some confusion about this point exists in the literature, and some published papers conflate these two numbers.

online at http://genetics.bwh.harvard.edu/cpd.

The remainder of this chapter is, at the time of this writing, in press in *Nature*[99]. It is reproduced here with permission. Supplementary material will be published online, but is included in this dissertation in appendix A. My primary contributions to this work are the design and implementation of the bulk analysis of multiple sequence alignments described in 2.2, the statistical model of CPD fixation described in 2.3, and the CPD predictor outlined in 2.5.

2.1 Background

Understanding the nature and prevalence of genetic interactions has the potential to inform the evolutionary forces that act on specific protein residues, protein complexes and, more broadly, the evolution of genomes. Some recent studies have reported that interactions are ubiquitous and contribute significantly to the evolutionary landscape[16,91,124,192],while others found that interactions are rare and that protein evolution can be modeled without taking them into account[39,81,130]. Even among those who agree that interactions are important, the architecture of these interactions remains unclear: some studies find distinct interactions between two or three sites[36,60,75], while others propose a complex interaction network, effectively responding to aggregate properties of the entire protein or the entire genome[31,91,193].

One practical utility of comparative genomics has been highlighted by our emerging appreciation of the large number of rare variants in humans and the difficulty in inferring their contribution to disease phenotypes[34]. To prioritize variants of interest, their frequency in control populations and their evolutionary conservation have become two prominent filters[179]. Conserved regions are considered more likely to be biologically important and intolerant of variation[4,11,122]; programs such as PolyPhen[2]and SIFT[165]have employed this principle to predict the functional effects of variants in both the research and clinical setting[7,34,101]. Although clearly useful, these strategies are constrained,

in part because they do not take into consideration the genomic context of the mutated allele. An allele can appear to be damaging in one sequence yet be neutral in an orthologous sequence of another species. This phenomenon, referred to as compensated pathogenic deviation (CPD), contributes an unknown, but potentially significant, number of false negatives to the evaluation of functional sites[101,112,115].

2.2 Prevalence of CPDs

To examine the prevalence of CPDs, and to identify such sites, we used comparative genomics on a genome-wide scale. A typical, non-CPD allele should cause the same phenotype in any orthologous sequence, regardless of genetic background. By contrast, when a variant that causes human genetic disease is found in a wild-type (wt) orthologous sequence, it is likely that the genetic background of that species exerts a compensatory effect on such a variant: it suppresses the phenotype otherwise caused by the variant, and thus protects the variant from negative selection[60,112,115,167]. Previous studies have used this insight to quantify the fraction of interspecies substitutions that are CPDs at approximately 10%, regardless of the distance between the two species[112,115,167]. Other studies have also reported estimates of the inverse value, namely the fraction of pathogenic variants that are present as CPDs in other species, ranging from 2%–18%, depending on the exact sequences and methodologies[30,55]. We set out to produce a new estimate of this value, taking advantage of the availability of large datasets of sequence variants and multiple sequence alignments. We collected two datasets of missense single-nucleotide variants (SNVs), annotated as either benign, or implicated in human disease, derived from two databases, one based on the literature ("HumVar")[2,22,139] and one based on reports from clinical genetic laboratories and investigator-initiated submissions ("ClinVar")[118]. Although the two databases are not fully independent, since some variants are present in both from the same source, the majority of pathogenic variants were listed in one or the other (Figure 2.1A).

Table 2.1: Fraction of likely pathogenic mutations in humans considered CPDs according to different filtering paradigms. Values represent the fraction of variants for which an alignment could be retrieved where the variant amino acid is present in an ortholog sequence; error ranges are Jeffreys 95% confidence intervals.

	Unfiltered Alignment (MultiZ)	High-Quality Alignment (MultiZ)	Mammalian Subset (MultiZ)	Multiple Species (MultiZ)	EPO Alignment
HumVar	12.0% ± 0.5%	11.5%±0.5%	6.7%±0.4%	6.1% ± 0.3%	7.5%±0.4%
ClinVar	10.2% ± 0.7%	9.9%±0.7%	5.6% ± 0.5%	4.7%±0.5%	6.5%±0.6%
HumVar + ClinVar	9.3% ± 1.0%	8.5% ± 1.0%	5.3% ± 0.8%	3.9%±0.7%	5.5% ± 0.9%
HumVar + ClinVar + ESP	7.5% ± 1.0%	7.0%±1.0%	3.8%±0.7%	3.0%±0.6%	4.0%±0.8%

In total, these datasets comprised 69,905 human missense mutations across 13,040 genes. We then compared this dataset to multiple sequence alignments of orthologous proteins from 100 vertebrate species[103]. As expected, we found the mutant residue for a large number of likely neutral human variants to be fixed in orthologs. However, the number of pathogenic missense variants found in orthologs (CPDs) was surprisingly high: 5.6% ± 0.5% of ClinVar and 6.7% ± 0.4% of HumVar were found in the alignment of mammalian species. For all vertebrates, these numbers increase to 10.2% ± 0.7% and 12.0% ± 0.5%, respectively, for a combined rate of 11.7% ± 0.4% (Table 2.1), although the alignments for non-mammalian vertebrates are not supported by synteny.

Mindful of the possibility that our allele set was contaminated with false pathogenic annotations[27,111], we applied a number of filtering steps with the goal of estimating the range of CPD incidence in humans. These steps included cross-referencing HumVar and ClinVar variant annotations with each other and with population frequency data[179], filtering based on alignment quality[3], using alternate alignment methodologies[57], and requiring presence in multiple species (Table 2.1; Supplementary Note). In addition to removing false positives, some of these filters did remove bona fide

recessive alleles (since we did not allow carriers), as well as disease variants with incomplete penetrance, even though this class of alleles is, by definition, sensitive to genomic context and thus likely to be affected by compensation[49,68]. Nevertheless, all filtering steps retained a substantial number of variants (Table A.1); post hoc manual evaluation and literature review yielded numerous robust examples of alleles that drive acute genetic disorders (Table A.2). Importantly, including only variants that pass all filtering steps and are detected in more than one vertebrate, we still predict that the minimum estimate of CPDs in human patients is 3% (Figure 2.1B). This is consistent with previous analyses, which have found that stringent filtering does not change the observed properties of CPDs significantly[55,167]. As a final test, we extracted *post hoc* likely pathogenic alleles from three different sources, each of which used independent means for assessing pathogenicity in acute pediatric disorders: the 183 non-synonymous pathogenic alleles reported recently in the Deciphering Developmental Disorders (DDD) study[56]; the 161 curated non-synonymous alleles from the genes that cause neuronal ceroid lipofuscinosis[113]and the 49 alleles implicated as recessive drivers of Bardet-Biedl syndrome that have also been annotated functionally *in vivo*[206]. Considering our stringent mammalian alignments only, the CPD rates were 9%, 5.5% and 4% respectively. Of note, when we only considered *de novo* alleles from DDD in the mammalian alignment exclusively and absent from control databases, the most conservative boundary for CPD incidence from this analysis remains at 3%. We performed several additional analyses to rule out other possible sources of bias (see Supplementary Note, Tables A.3, A.4, A.5); these analyses were also consistent with previously reported properties of CPDs[8,55].

2.3 Structure of Genetic Interactions

We next turned to the question of the structure of the genetic interactions underlying such sites. In broad terms, there are two possibilities: suppression of the disease phenotype may be the result of a

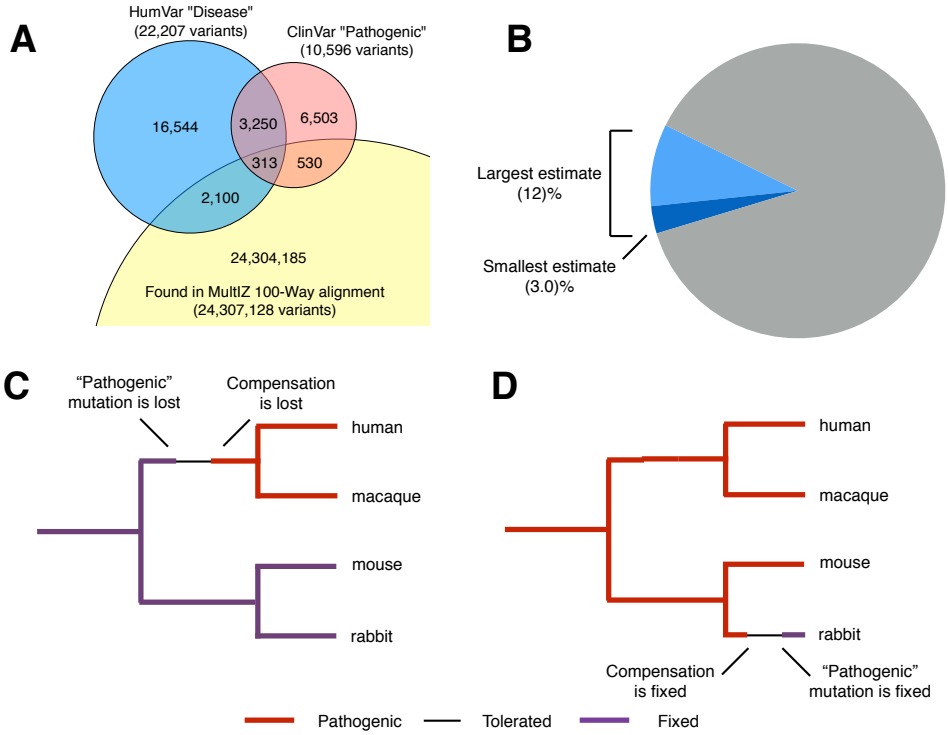


Figure 2.1: Distribution of variants found in sequence alignments. A) Venn diagram showing sizes and overlap of the ClinVar and HumVar datasets of human disease variants, and how many are found in the multiple sequence alignment. B) Illustration of the estimated number of human disease variants found in the alignment. The smallest estimate (3.0%, dark blue) comes from using the intersection of both variant datasets, requiring the variant be absent from 6,503 human exomes, and filtering out alignments with low quality scores. With any methodology, at least 88% of human disease variants (grey) are not found in the alignment. C, D) Illustration of potential mechanisms for the occurrence of compensated pathogenic mutations in evolution. Branches where the variant is fixed are purple; branches where the variant is pathogenic are red. In panel C, the variant is present neutrally in an ancestral population, but is lost in the primate branch. Subsequent substitutions cause the ancestral allele to become pathogenic. In panel D, the variant is pathogenic in the ancestral population, but mutations in a non-human branch cause it to become tolerated, and it arises later by mutation and becomes fixed.

small number of discrete compensatory substitutions; or suppression may be caused by global shift in the properties of the gene, or even of the whole genome, caused by a large number of discrete substitutions that, individually, have small effects. These two models have different ramifications for detecting the mechanism(s) that drive CPDs. If only a small number of substitutions are involved, we can attempt to identify individual *cis*-compensatory substitutions by studying the co-evolution of the CPD residue with other residues within the target protein. By contrast, this kind of experiment is virtually impossible if multiple genes or the entire genome is involved.

The difference between these two models should be visible in the distribution of CPDs among orthologous sequences. In the course of evolution, variants arise stochastically through a Poisson process, where the expected amount of evolutionary time required to produce a given substitution is distributed exponentially. For a CPD, however, the distribution should be different. This is because the presence of a CPD also mandates the presence of all the compensatory substitutions necessary for the CPD to be rendered neutral. As such, the expected evolutionary time required to produce a CPD is the sum of the times required to produce each necessary compensatory substitution, followed by the time required to produce the CPD itself.

Previous studies have proposed different processes by which CPDs can arise. The most plausible option is a neutral mechanism, where the compensatory substitutions are themselves neutral and arise and fix neutrally before the pathogenic substitution appears (Figure 2.1C, D; Figure 2.2A). In this case, the time required for each substitution to arise is given by an exponential distribution, and the time for all compensatory sites to arise is approximated by the convolution of multiple exponential distributions (or a gamma distribution, in the special case where all these exponential distributions are identical). The number of exponential distributions included in the convolution corresponds to one plus the number of compensatory substitutions required, and it can be inferred from the shape of the distribution (Figure 2.2B).

Although the actual evolutionary time separating two sequences is not observable directly, we

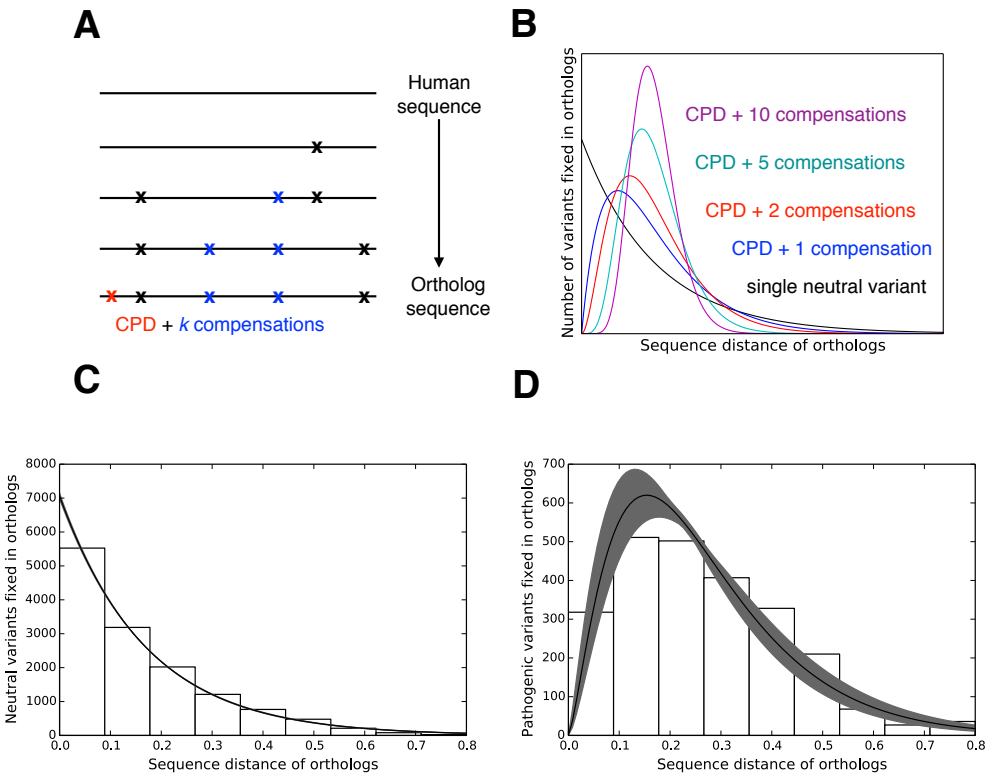


Figure 2.2: Relationship between variants and evolutionary distance. A) Model for fixation of CPDs. Neutral changes (Xs) arise neutrally over evolutionary time. Some of these (blue) compensate for alleles that would otherwise be pathogenic. Once k of these compensatory changes have fixed, the CPD (red) can fix neutrally. B) The relationship between evolutionary distance and the number of variants observed in the alignment is expected to be different for individual benign variants (black) and pathogenic variants with different numbers of compensations (blue: one, red: two, cyan: five, magenta: ten). C, D) Observed distribution of missense variants annotated as neutral (C) or pathogenic (D) in the HumVar dataset and present in vertebrate orthologs (bars), with maximum likelihood fits (black lines) and 95% confidence bands (gray shading). Panel D corresponds to a fitted value for k of 1.44 ± 0.07.

can approximate it using sequence distance (one minus sequence identity).We therefore plotted the number of missense variants observed as a function of sequence distance for neutral variants and for CPDs. Qualitatively, the shapes of both distributions match theoretical expectations. The two distributions are remarkably distinct from each other ($p = 1.6 \times 10^{-68}$, Kolmogorov-Smirnov 2-sample test; see Tables A.6, A.7), providing additional evidence that variants annotated as neutral and variants annotated as pathogenic represent biologically distinct classes. Additionally, the observed distribution of CPDs is weighted toward shorter evolutionary distances, as expected if most CPDs require only a small number of individual compensatory substitutions, and does not resemble the normal distribution expected if most CPDs require many individual compensatory substitutions (Figure 2.2B, D). To obtain a more precise estimate of the number of compensatory substitutions, we used maximum likelihood to fit several versions of the convolution-of-exponentials model with several different combinations of variant datasets and alignment strategies (Figure 2.2C, D; see methods and Tables A.6, A.7). Most versions of the model fit best as the convolution of approximately two exponential distributions, supporting a mechanism whereby the majority of CPDs are compensated by simple pairwise interactions. Additionally, most models reported similar rates of evolution for neutral variants, CPDs, and compensatory variants, suggesting that the target size for compensatory changes is small. We repeated these analyses with multiple different variant datasets and alignment strategies, finding similar results each time (Figure A.1, Table A.8).

These analyses predict that most CPDs could be rescued by one large-effect compensatory substitution. We tested this prediction experimentally. We posited that each vertebrate sequence that includes a CPD should also include its cis-compensatory allele. Therefore, every amino acid difference between the human sequence and the sequence of the ortholog(s) containing a CPD is a candidate compensatory substitution. Given the practical constraints of examining all possible compensatory substitutions in macromolecular complexes, we focused on substitutions within the same gene as the CPD. This restriction is supported by previous observations, wherein 90% of compensatory

substitutions found in eukaryotes lie within the same gene as the pathogenic substitution for which they compensate[156].

2.4 *In Vivo* Validation of CPDs

Scanning our list of candidate CPDs, we noted two alleles in genes involved in ciliopathies: a p.N165H encoding change in *BBS4* and a p.R937L variant in *RPGRIP1L*, which contribute pathogenic alleles to Bardet-Biedl syndrome and Meckel-Gruber syndrome, respectively[105,106]. These alleles were prioritized to test our predictions for several reasons: a) those disorders have a severe effect on reproductive fitness; b) previous studies have established loss-of-function zebrafish phenotypes rescuable by human mRNA for both genes[106,206]; c) in vivo complementation with each allele has indicated both of them to be deleterious to the function of the human protein[106,206]; and d) we observed multiple species with the human mutant allele fixed: six species for *BBS4* 165H and four species for *RPGRIP1L* 937L (Figure 2.3A, B); for this reason, both alleles were predicted to be benign by several computational methods (PolyPhen-2, SIFT, MutationAssessor).

Comparative genomic analysis of *BBS4* and *RPGRIP1L* identified nine candidate sites in human *BBS4* and 32 candidate sites in human *RPGRIP1L* (519 aa and 1315 aa of target sequence respectively; Table A.9) that co-evolved with the deleterious allele. To test each site, we took advantage of the established convergent extension (CE) defects induced by morpholino (MO)-mediated suppression of bbs4 or rpgrip1l in zebrafish embryos[106,206]. Consistent with previous observations, suppression of bbs4 or rpgrip1l induced CE defects in 80% and 50% of embryos respectively (50-100 embryos, performed in triplicate, blind scoring; Figure 2.3C–E). Co-injection of MO with human wt mRNA rescued this phenotype, whereas injection with human mutant mRNA showed no improvement (Figure 2.3D, E). We next tested the entire candidate complementing allelic series for each gene. For *BBS4*, the introduction of two of the nine candidate residues in cis with the 165H-encoding mRNA

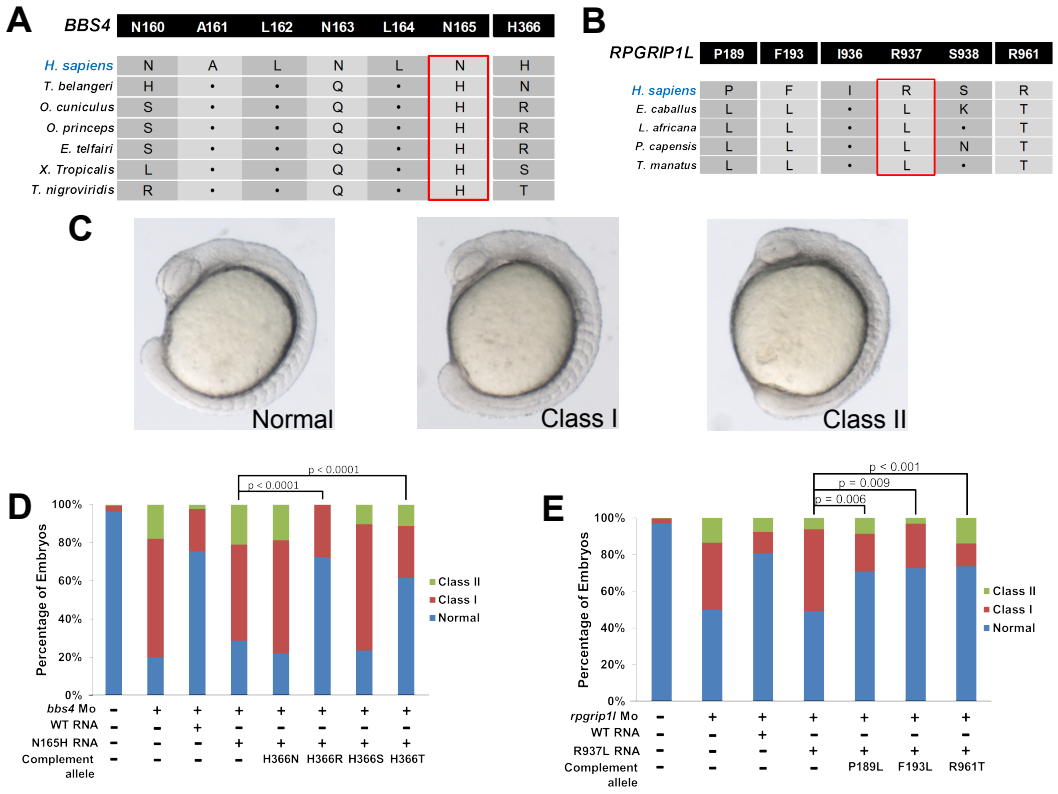


Figure 2.3: Compensatory mutations rescue pathogenic alleles in *BBS4* and *RPGRIP1L*. A) The pathogenic *BBS4* 165H allele is fixed in six species. Secondary sites 160, 163, and 366 are possible sites of complementation. B) The pathogenic *RPGRIP1L* 937L allele is fixed in four species. The 189L, 193L, and 961T alleles are present in all four species. C) Examples of zebrafish convergent extension phenotypic groups. D) Human RNA containing both the *BBS4* 165H mutation and either the 366R or 366T mutation can rescue the morphant phenotype, whereas RNA containing the 165H mutation alone cannot. E) Mutation of 189L, 193L, or 961T, in the background of 937L *RPGRIP1L* mRNA, is able to rescue the loss of function observed in 937L RNA. Significance determined by χ^2 test; see Table A.9 for raw data.

ameliorated the phenotype in a manner indistinguishable from wt mRNA. Strikingly, both complementing alleles affected the same amino acid and were specific with regard to the compensatory changes: the 165H/366N or the 165H/366S behaved as functionally null, whereas 165H/366R was indistinguishable from wt; 165H/366T converted the functional null to a hypomorph (Figure 2.3D; Figure A.2A).

We observed a similar pattern for *RPGRIP1L*, despite the larger number of candidate complementing sites. Testing each of the 32 candidates identified three complementing events, two of which map to the same region: 937L/189L, 937L/193L and 937L/961T (Figure 2.3E; Figure A.2B). Crucially, testing each complementing allele individually on wt background human mRNA showed them to be either extremely mild for *BBS4* or benign for *RPGRIP1L* (Table A.9). Finally, comparative genomic analysis of these sites showed that these data could explain the tolerance of the *RPGRIP1L* 937L-encoding allele in all four species and of the *BBS4* 165H-encoding allele in four of the six species (Figure 2.3A, B).

The above analysis is limited by its retrospective nature. We therefore tested the usefulness of our model in the context of *ab initio* gene discovery. We have initiated recently a whole exome sequencing (WES) and functional testing paradigm to accelerate gene discovery and diagnosis in young children with suspected genetic disease: The Duke Task Force for Neonatal Genomics (TFNG). Patients who display anatomical phenotypes amenable to functional modeling in zebrafish are evaluated by trio-based WES and have candidate alleles tested systematically by *in vivo* complementation[104].

As part of the TFNG, we enrolled a 17-month-old female with an undiagnosed neuroanatomical condition hallmarked by microcephaly (Figure 2.4A). We filtered WES data to retain non-synonymous and splice variants with a minor allele frequency below 1%, and we conducted a proband-centric trio analysis. This strategy yielded four candidates: de novo missense changes in two genes, *BTG2* and *NOS2*; and autosomal recessive missense variants in *TTN* and *LAMA1*. Testing of an

unaffected female sibling excluded *LAMA1*; and we considered *TTN*, a known autosomal dominant cardiomyopathy locus[133], to be an unlikely driver.

To investigate the allele pathogenicity of the *BTG2* (p.V141M) and *NOS2* (p.P795A) changes, we studied *btg2* and *nos2* in zebrafish embryos. Reciprocal BLAST identified a single zebrafish *btg2* ortholog and two zebrafish *nos2* orthologs. We injected splice-blocking MO or translational-blocking MO (sb-MO and tb-MO; Figure A.3) into wt zebrafish embryos (3 ng; n=80 embryos/injection). Given the microcephaly in the patient, we scored for head size defects at 4 days post-fertilization (dpf) by measuring (blind to injection) the anterior-posterior distance between the forebrain and the hindbrain (Figure 2.4B). For nos2a/b MO-injected embryos, we saw no appreciable differences at the highest dose injected (8 ng for nos2a/b sb-MOs; Table A.10). In contrast, we found a significant reduction of anterior structures in btg2 morphants ($p<0.0001$; Figure 2.4B, C). Co-injection of wt human *BTG2* mRNA with tb-MO resulted in significant rescue ($p<0.0001$; Figure 2.4C). In contrast, upon blind scoring of embryos injected with either wt or 141M-encoding mRNA (in triplicate), we found that 141M was significantly worse at rescue than wt ($p<0.0001$; Figure 2.4C).

BTG2 is a regulator of cell cycle checkpoint in neuronal cells at the G1 to S phase[138] and is strikingly intolerant to variation in humans (Exome Variant Server, EVS). To gain insight into the underlying cause of microcephaly in btg2 morphants, and to test the pathogenicity of the discovered allele by a different assay, we performed antibody staining at 2 dpf (a time prior to the manifestation of microcephaly). We marked post-mitotic neurons in the forebrain with HuC/D, and we scored (blind, in triplicate) based on an established paradigm[13]. *btg2* morphants displayed a significant decrease in HuC/D staining (97% vs. 50% of embryos with normal expression/localization; controls versus MOs, respectively; $p<0.0001$; Figure 2.4D, E). This defect was rescued with WT *BTG2* mRNA ($p<0.05$); but could not be ameliorated by 141M-encoded mRNA co-injection (Figure 2.4D, E). Importantly, co-injection of *btg2* tb-MO with two rare control EVS alleles (p.A126S and p.R145Q, each present in 1/13,006 and 1/12,994 chromosomes, respectively) resulted in rescue of

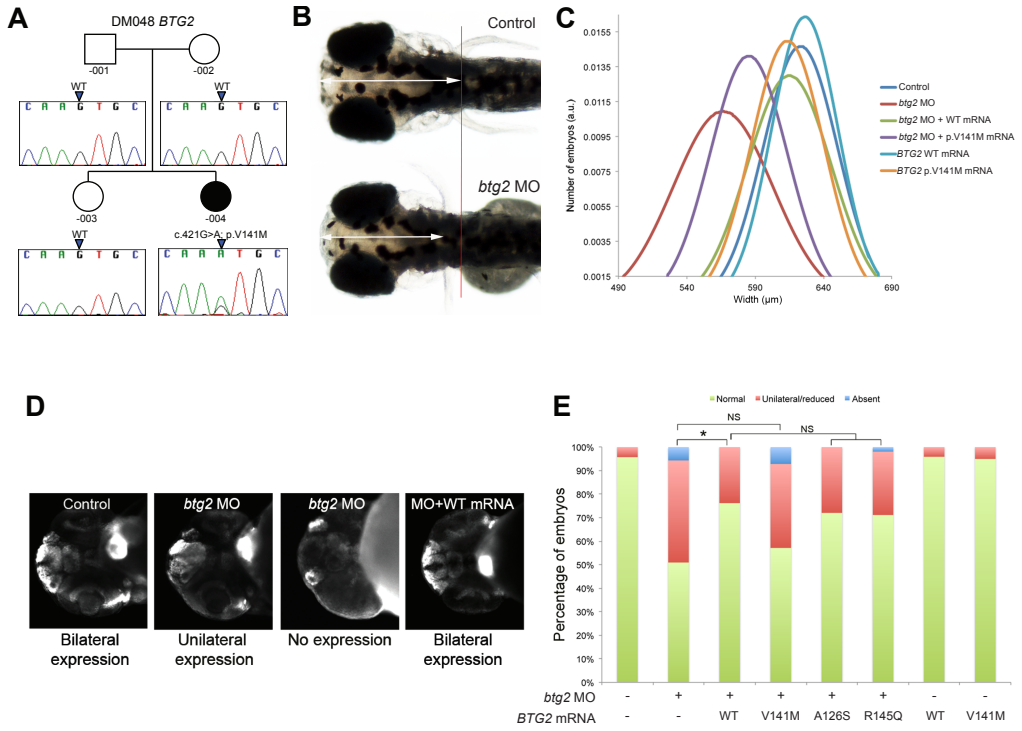
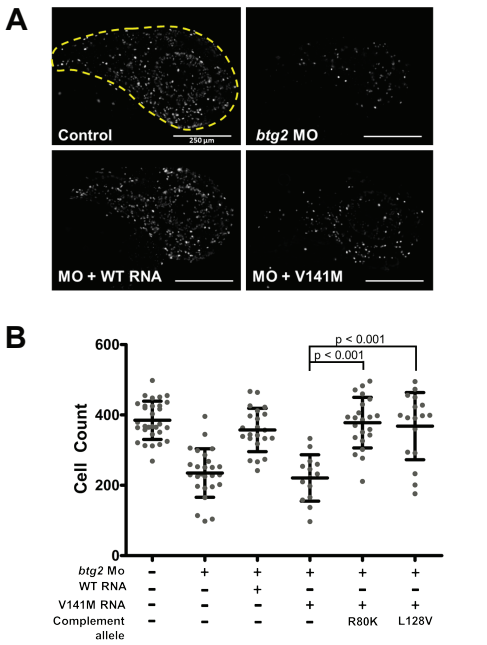


Figure 2.4: A *de novo* BTG2 p.V141M encoding allele causes microcephaly. A) Pedigree of DM048. Chromatograms show a *de novo* c.421G>A change in the index case. B) Suppression of *btg2* leads to head size defects. Dorsal view of uninjected (UI) control and *btg2* MO-injected zebrafish embryos at 4 dpf. White arrows show the distance measured from forebrain to hindbrain. Red line shows the protrusion of the pectoral fins in UI controls (for comparison with *btg2* morphant). C) Distribution of head size measurements at 4 dpf (Table A.10; white arrows in C), a.u., arbitrary units; $n = 49$–79 embryos per injection. D) Suppression of *btg2* leads to a decrease of HuC/D levels at 2 dpf. Representative ventral images of control, *btg2* morphants (images show unilateral or absence of HuC/D expression), and a rescued embryo injected with a *btg2* MO plus human *BTG2* wt mRNA. E) Percentage of embryos with normal, bilateral HuC/D protein levels in the anterior forebrain or decreased/unilateral HuC/D protein levels in embryos injected with *btg2* MOs alone or MOs plus human *BTG2* WT or variant mRNAs (p.V141M, index case; p.A126S and p.R145Q, control alleles). $*$: $p < 0.05$ (two-tailed t-test comparisons between MO-injected and rescued embryos; $n = 38$–78 embryos per injection batch).

HuC/D staining comparable to wt mRNA, providing evidence for assay specificity (Figure 2.4E). As a third test, we stained whole embryos with a phospho-histone H3 (PH3) antibody that marks proliferating cells. We counted the number of positive cells in a defined anterior region of a minimum of 10 embryos injected with each cocktail (scored blind). We saw a significant reduction in cell proliferation in the heads of 2 dpf *btg2* morphants (average of 235 PH3-positive cells/embryo head in morphants vs 387 in controls, $p<0.0001$); this defect was likewise rescued by co-injection of wt mRNA, while 141M mutant rescue was indistinguishable from *btg2* tb-MO alone ($p = 0.38$; Figure 2.5A, B). Combined, all three assays indicated that *BTG2* p.V141M is pathogenic and that haploinsufficiency of this gene likely contributes to the microcephaly phenotype of the proband.

Despite all our functional and genetic data for p.V141M, this allele was predicted to be benign by each of Polyphen-2, SIFT, and MutationAssessor. The likely reason is that, with the exception of primates, most other species with a *BTG2* ortholog encode Met at the orthologous position (Figure 2.5C). These data suggested that V141 might represent a CPD site in primates that branched from the ancestral methionine residue. To test this possibility, we searched for *BTG2* sites that co-evolved with 141M. We identified nine such sites (Table A.11), which we mutagenized into the human construct bearing the 141M-encoding allele. We then injected embryos with btg2 MO; MO+wt human *BTG2* mRNA; MO+ 141M-encoding mRNA; or MO+ 141M in *cis* with one of the nine candidate complementing alleles. Seven of the alleles had no effect on this activity (Table A.11). However, R80K- or L128V-encoded mRNA on the 141M backbone rescued the number of PH3-positive cells to wt levels (Figure 2.5B; Figure A.3C); both alleles were benign on their own (Table A.11). Taken together, our data indicated that 141M is deleterious to protein function in the human background, but the protection to this residue afforded by either Lys 80 or by Val 128 can explain more than 90% (54/59) of species encoding 141M (Figure 2.5C).



C

BTG2	R80	L128	Q140	V141	L142
H. sapiens	R	L	Q	V	L
P. troglodytes	•	•	•	•	•
G. gorilla	•	•	•	•	•
M. musculus	K	V	•	M	M
R. norvegicus	K	V	•	M	M
H. glaber	•	V	•	M	M
S. domesticus	K	V	•	M	M
B. primigenius	K	V	•	M	M
E. ferus caballus	K	V	•	M	M
F. catus	K	V	•	M	M
C. lupus familiaris	K	V	•	M	M
D. novemcinctus	K	V	•	M	M
G. gallus	K	P	•	M	M

Figure 2.5: 80K and 128V mutations rescue the *BTG2* V141M allele. A) Two day-post-fertilization zebrafish embryos stained for phospho-histone H3. Human RNA containing the V141M mutation is unable to rescue the reduction of proliferation observed upon knockdown of *btg2*. B) Quantification of pH3: human RNA with mutations V141M and either R80K or L128V can rescue MO knockdown of endogenous *btg2*. Error bars represent standard deviation. C) The 141M allele is fixed in 58/87 species besides primates, examples displayed here. See Table A.11 for raw data.

2.5 Discussion

The ability to test allele function in a biologically relevant context is improving. Nonetheless, these methods cannot yet be employed at the genome or population level. To improve the scalability of detecting CPDs, we used our model of CPD evolution to develop a computational predictor for distinguishing variants that are unlikely to be CPDs from those that might be CPDs, and to identify candidate compensations to aid in experimental design (`http://genetics.bwh.harvard.edu/cpd/`). Initial testing of this tool intimated high negative predictive values but modest positive predictive values, likely due to the dearth of known CPDs (see Supplementary note).

These results contrast some previous studies that claim that epistasis is ubiquitous[16,36,124];or that it is practically nonexistent[6,130]; or that it is commonly of higher order[31,193]. The most likely explanation for this discrepancy is that most such studies have examined different kinds of variation and different kinds of traits. For example, studies on the evolution of genetic incompatibilities rely on assumptions of high mutation rate and weak negative selection, assumptions that generally do not hold for the case of pathogenic missense variation[36,59]. The difference with the studies suggesting higher-order *cis*-interactions may have to do with the scale of evolutionary time our analyses probes: the span of hundreds of millions of years of evolution represented by the vertebrate alignment may not be long enough to reveal higher-order combinations of nonsynonymous SNVs. Indeed, using neutral SNVs from the HumVar dataset as a control, we estimate the vertebrate alignment has explored 12% of pairwise interactions between SNVs, compared to 0.6% of three-way interactions between SNVs. It is possible that higher-order interactions are common, but are not detectable without a deeper alignment.

Finally, in the backdrop of accelerated use of genome editing to model human pathogenic mutations in a variety of model organisms, our data highlight the critical need to not only pair computational predictions with functional studies, but to also evaluate the effect of human mutations in the

context of the human sequence.

Why ask about my birth?
Like the generations of leaves, the lives of mortal men.
Now the wind scatters the old leaves across the earth,
now the living timber bursts with the new buds
and spring comes round again. And so with men:
as one generation comes to life, another dies away.
But about my birth, if you'd like to learn it well,
first to last—though many people know it—
here's my story.

Homer, *Iliad* book 6[83]

3

Parametric rate estimation in proteins

The Bayesian rate estimator score included in the HCM predictor reported in chapter 1 turns out to be a very useful feature. Unlike the other scores developed in that chapter, it is not dependent on the specific genes under study and their biochemical or biophysical properties; it is in principle applicable to the entire genome. Furthermore, it appears to add information to the unmodified PolyPhen-2 score despite being ostensibly based on the same information (the multiple

sequence alignment). If it could be plausibly precomputed or made faster, it would be possible to include this score in the general-purpose version of PolyPhen. With this in mind, I set out to implement a faster score that incorporates the same parametric model as this Bayesian rate estimator score.

I investigated several possibilities for this score. One class of rejected estimators attempted to generate an equilibrium distribution of amino acid frequencies π, the same quantity PolyPhen and SIFT estimate. Another class allowed for shifting amino acid preferences throughout the tree, attempting to incorporate my model of multisite interactions from chapter 2. Ultimately all of these models appeared to have too many parameters for the number of sequences we have in our alignment. This problem may be exacerbated by the flatness of the likelihood landscape, as alluded to in the introduction. It is possible that these models may become tractable once our alignments contain thousands of vertebrate sequences rather than hundreds. Ultimately the score that worked best was a simple rate, very similar to the Bayesian rate estimator score reported in chapter 1. It is also fairly similar to to several existing parametric methods, though those methods use nucleotide sequences rather than amino acid sequences, a fact which may contribute to performance[33,41,74,154].

The remainder of this chapter is a manuscript that is currently under review for publication. The cross-validation procedure was designed and implemented by Ivan A. Adzhubey; the remainder of the text is my work, with advice and support from Professor Sunyaev.

3.1 Background

Estimating the level of selective constraint on a sequence is one of the most common tasks in computational genetics. In addition to being important to the history and dynamics of evolution, selective constraint can be used as a proxy for biological importance or strength of phenotypic effect[4,93,145,170,207]. In this context, estimates of selective constraint are often used for such purposes

as selecting variants that are likely to be causative of rare Mendelian disorders[7,34], developing animal models of diseases[62], and identifying possible binding sites for transcription factors[79,195], among others. There are many existing methods designed for these purposes, including complete phylogenetic analysis packages that are capable of producing rate estimates, such as PAML[201], MrBayes[161], PLEX[42], and HyPhy[140,155], as well as tools that have estimation of constraint as their primary purpose, like phyloP[154], GERP++[41], LRT[33], and fitCons[74].

Many phylogenetic analysis programs include estimations of evolutionary rate as part of their analysis. In particular, it is common for these tools to use a parametric model of evolution with multiple rate categories or a distribution of rates, with different sites evolving under different levels of constraint, or, in some cases, under different levels of positive selection. These rate categories are inferred along with the tree structure, and provide information about the dynamics of evolution over an entire gene or series of genes. These tools are extremely versatile, supporting a wide variety of models of evolution and methods of phylogenetic inference. However, for the particular purpose of rate estimation they are fairly coarse-grained, with one rate category typically containing many nucleotides or codons, and cannot generally be used to evaluate the level of constraint at a particular site.

Methods like phyloP, GERP++, LRT, and fitCons are specifically designed to address the problem of assigning precise levels of constraint to particular sites. These methods vary in their specific models and implementations, but they generally use a similar approach. Given a multiple sequence alignment and a precomputed tree describing the relationships between the sequences in the alignment, these methods use parametric models of evolution to estimate the likelihood that a particular site is evolving under constraint. Each of these methods uses a slightly different statistical formulation of the problem and computes its final score differently, but each effectively uses a likelihood model to estimate the rate of evolution, and then scores whether that rate is consistent with neutral evolution or implies constraint. These methods are typically used to detect patterns of conservation

in sequences, including detection of blocks of conserved sequence. Some methods, such as fitCons, also incorporate other sources of information to distinguish functional regions, including patterns of DNAse hypersensitivity and histone modifications.

These methods typically use nucleotide sequences only, and therefore may fail to capture some biochemical information about the protein sequence. Those that do use amino acid sequences still explicitly model substitutions at the nucleotide level, using a codon model. The main reason for this is the tractability of the model: where a nucleotide-level model of evolution must supply or infer a 4×4 substitution matrix (one row and column for each of the four standard nucleotides), or 12 independent rate parameters, a full amino acid model requires a 20×20 matrix (one row and column for each of the twenty standard amino acids), or 380 independent rate parameters. The extremely high dimensionality of the amino acid model makes inference of the model parameters unfeasible for the typical use cases of these tools, which involve a single site observed in fewer than 100 taxa.

While a full parametric treatment of an amino acid substitution model might not be feasible, it is in principle possible to estimate certain relevant features of such a model. One useful feature is the stationary distribution of amino acid frequencies at a particular site, which in principle represents the profile of amino acids tolerated at that site[101]. Several methods exist to estimate this quantity from a multiple sequence alignment, including SIFT[142] and PSIC[171], which may be more familiar as a component of ensemble predictors of variant effect such as PolyPhen-2[2] or SNAP[18]. These methods forego the parametric model, instead using the observed distribution of amino acids produced by evolution within a given phylogeny to estimate the stationary distribution of amino acids. They consistently outperform parametric rate estimation methods like phyloP and GERP++ at the task of separating neutral from deleterious mutations[52,108], demonstrating that the stationary distribution of amino acids contains useful information about evolutionary constraint that is not available to parametric models.

Here, we present a new parametric method for estimating evolutionary rate, BARE (Bayesian Amino-Acid Rate Estimator). Unlike existing methods, BARE uses a pure amino acid substitution model. We use an estimate of the stationary distribution of amino acids produced by PSIC to constrain the model, allowing the parameters to be efficiently inferred at a single site with fewer than 100 taxa. This method outperforms existing parametric methods, as well as providing additional information apparently not captured by the raw PSIC profile.

3.2 Methods

3.2.1 Model

Our method is based on a standard parametric model of protein evolution from the generalized time-reversible (GTR) class of models, which models evolution as a branching Markov process with a constant transition rate matrix Q[92]. The rate matrix Q represents the instantaneous rate of transition between states: Q_{ij} for $i \neq j$ represents the rate of transition from state i to state j, with the diagonal elements Q_{ii} given by the requirement that the columns of the matrix sum to zero, i.e. that

$$\sum_j Q_{ji} = 0 \tag{3.1}$$

The rate matrix Q is used to compute the transition probability matrix $P(t)$, where $P_{ij}(t)$ represents the probability of observing state j at time $t_0 + t$ conditional on observing state i at time t_0. $P(t)$ is given by

$$P(t) = e^{Qt} \tag{3.2}$$

where e^{Qt} represents the matrix exponential.

The elements of the transition rate matrix Q_{ij} are given by

$$Q_{ij} = \begin{cases} \mu \vartheta_{ij} \pi_j & \text{if } i \neq j \\ -\sum_{k \neq i} Q_{ki} & \text{if } i = j \end{cases} \tag{3.3}$$

where π is the vector of stationary amino acid frequencies, ϑ is a rate matrix representing the exchangeability of amino acid states independent of their frequencies, and μ is a scale factor applied to the entire matrix. This formulation guarantees that the stationary distribution of amino acids π and the rate matrix Q obey the stationarity criterion

$$\pi \cdot Q = 0 \tag{3.4}$$

In other words, the net transition rate is zero when the distribution of amino acids matches the stationary distribution π.

This separation of the stationary state frequencies π from the exchangeability parameters ϑ makes it straightforward to constrain this model by choosing a fixed exchangeability matrix ϑ, allowing only the vector π and the scale factor μ to vary freely[20,92]. This produces a model with 20 free parameters, representing 19 independent stationary-state amino acid frequencies π_i and the scale factor μ. The 20th amino acid frequency is given by the normalization requirement,

$$\sum_i \pi_i = 1 \tag{3.5}$$

3.2.2 Priors and Constraints

We used a ϑ matrix derived from the empirical substitution matrix of Jones, Taylor, and Thornton[98]. We fix π for each individual site using PSIC[171], which estimates π based on a multiple se-

quence alignment of amino acid sequences. This is the essential novelty of our method, and it allows us to evaluate rapidly a parametric model for each amino acid site independently using the full alphabet of amino acids. Fixing ϑ and π leaves a single free parameter: the scale parameter μ, which represents the overall rate of evolution at the site. We applied this model to the phylogenetic tree provided by the MultiZ orthologous alignments of 100 vertebrate sequences, available on the UCSC genome browser[103]. We populated the tree with amino acid identities from a single site of the alignment, and computed a maximum *a posteriori* estimate of the single paramter μ. As a prior for μ, we used an uninformative gamma prior with parameters $\alpha = \beta = 1$.

3.2.3 Validation

There is no direct way to test the predictions of the method, because there are no known "true" values for the rate of evolution. One commonly used proxy is the task of separating known benign variants from known pathogenic variants, as sites containing pathogenic variation are more likely to be under evolutionary constraint. This task is also important for several of the most common and important use cases of estimates of variant effect, which include prediction of variant effect and prioritization of variants. We tested our method using the HumVar dataset of variant annotations[22], derived from amino acid variant annotations listed in the UniProt database (`http://www.uniprot.org/docs/humsavar`). We compared the performance to phyloP, GERP++, and LRT scores extracted from the dbNSFP database of variant annotations[123]; PSIC scores representing the log likelihood of the wild-type amino acid ("Score1") and the log likelihood ratio of the wild-type and variant amino acids ("dScore"); and the full ensemble of 11 features used by the current version of PolyPhen-2 (Figure 3.1). All scores were used as features of Naive Bayes classifiers with discretization, implemented in Weka (`http://www.cs.waikato.ac.nz/ml/weka/`), and trained and scored in 5-fold cross-validation, stratified so that each gene was only included in a single fold.

3.2.4 Availability

The method is implemented in C++ using the Bio++ library (http://biopp.univ-montp2.fr/), and is freely available under the GPL license at http://bwh.harvard.edu/bare.

3.3 Discussion

3.3.1 Performance

As a proxy for the accuracy of our method at predicting the level of evolutionary constraint, we measured its ability to distinguish known neutral variants from those known to cause disease in humans (Figure 3.1). Using our cross-validation procedure (see section 3.2.3), we find that our method outperforms existing parametric methods at this task. Our method's performance was comparable to but worse than that of PSIC, a non-parametric method that does incorporate amino acid identities. However, when combined with PSIC in an ensemble classifier, our method improves its performance noticeably, and to a larger degree than other tree-aware methods.

3.3.2 Precomputed Dataset

In our C++ implementation, the method is fast enough that we can precompute the rate score for all coding positions in all known transcripts of human protein-coding genes. We used this dataset to compare our rate score to precomputed scores from other methods stored in the dbNSFP database, excluding fourfold synonymous sites[123]. Our score shows reasonably good but not perfect correlation with other scores including phyloP (Spearman rho = -0.65), GERP++ (Spearman rho = -0.51), and LRT (Spearman rho = 0.68). Most discrepancies between the different methods are probably accounted for by the fact that BARE is aware of the chemical properties of different amino acids, while the other methods are not. Partitioning the dataset based on the amino acid type found at

each position confirms the importance of amino acid type: both BARE and the other methods reflect differences in conservation patterns between different amino acid types, as expected[98], but BARE shows remarkably different patterns from the other methods (Figure 3.2A–B). Different amino acid types also show a wide range of correlations, with Spearman rho values for specific types of amino acids ranging from approximately 0.5 to 0.8 (Figure 3.2C). Measuring prediction performance on this stratified dataset (Table 3.1) indicates that, although BARE has better performance overall, BARE does not necessarily have the advantage in all cases where BARE and phyloP disagree. For example, amino acid aware methods appear to perform consistently better in amino acids with charged side chains (aspartic acid, glutamic acid, arginine, and lysine), while nucleotide-only methods in amino acids with aromatic rings (histidine, phenylalanine, tryptophan, and tyrosine).

3.4 Conclusion

We have presented BARE, a new tool to score the evolutionary constraint on protein coding positions. Mathematically, this method is similar to the computations performed by the phyloP[154], GERP++[41], and LRT[33] methods to compute a "neutral rate" of evolution at each site, with the difference that BARE explicitly accounts for biochemical properties of amino acid sequences and is thus potentially more informative about coding sequences. Explicitly modeling amino acid sequences in this way has a significant impact on this rate calculation, especially for amino acids with important biochemical properties like aromatic rings or charged side chains.

Without a reliable way of measuring the "true" rate, it is impossible to say which estimator is more accurate, but we argue that modeling amino acid sequences is essential to creating an accurate map of constraint in protein-coding sequences. One indication that the BARE estimator performs well is that it more accurately separates pathogenic missense variants from neutral missense variants.

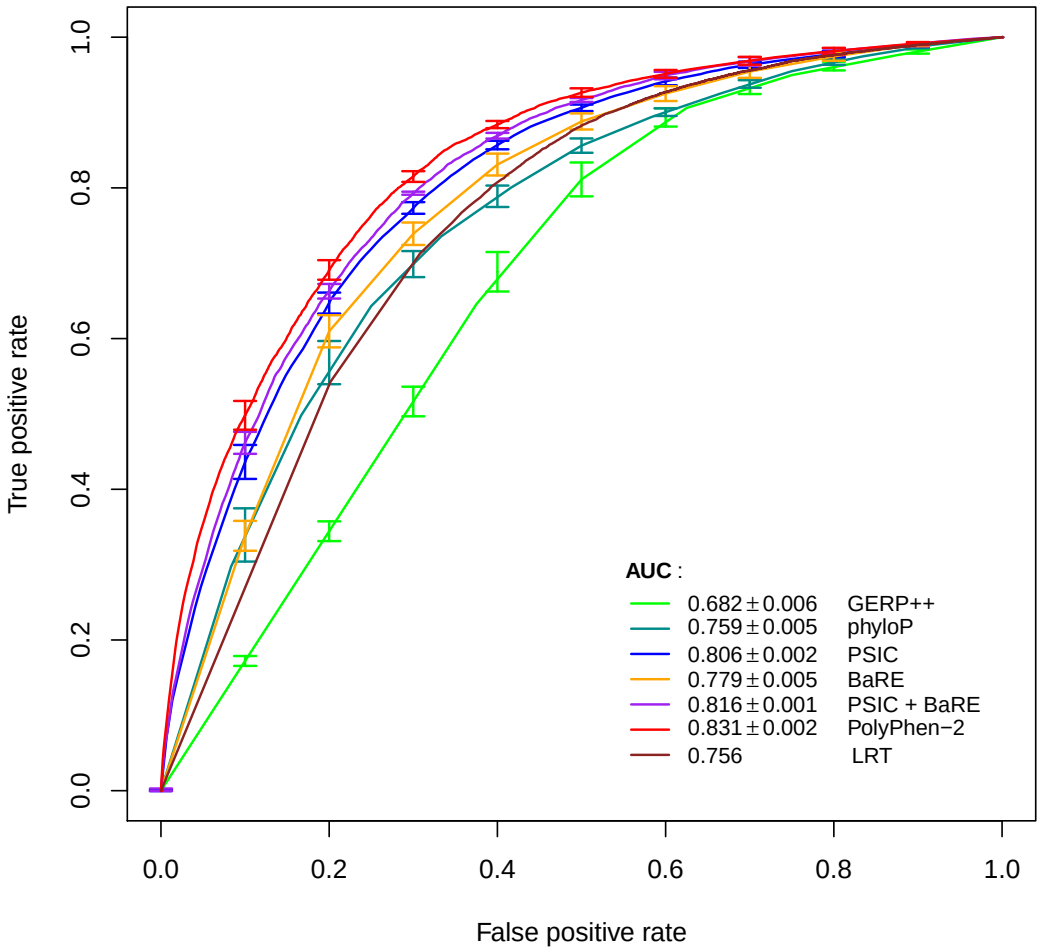


Figure 3.1: ROC analysis for BARE compared with other methods. "GERP++," "phyloP," and "BARE" represent single-feature Naive Bayes classifiers. "PSIC" represents a two-feature Naive Bayes classifier, using wild-type PSIC score (Score1) and difference in PSIC scores (dScore). "PSIC + BARE" represents a three-feature Naive Bayes classifier combining the two PSIC scores and BARE. "PolyPhen-2" represents the standard 11 features of PolyPhen-2, i.e. the current complete PolyPhen-2 classifier. LRT is a raw score taken from the dbNSFP database. All classifiers except LRT are averages from 5-fold cross-validation, with standard error ranges.

Table 3.1: Classification accuracy of Naive Bayes classifiers (testing-on-training) for different amino acid types. Amino acid aware methods and nucleotide-based methods show different profiles of accuracies depending on amino acid types, demonstrating that an amino acid level substitution model does make a significant difference in these rate estimates. Values marked with $*$ are statistically significant departures from the overall value (20-sample test for equality of proportions without continuity correction, with Bonferroni multiple test correction).

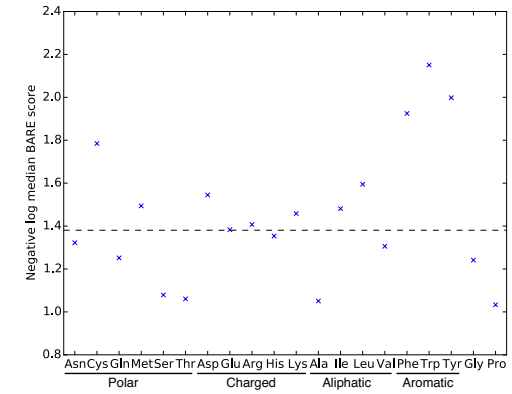
Amino Acid	GERP++	phyloP	PSIC	BARE	PSIC + BARE
Overall	0.66	0.70	0.74	0.73	0.75
Ala	0.67	0.74*	0.77	0.75	0.78*
Cys	0.77*	0.77*	0.82*	0.81*	0.82*
Asp	0.63	0.68	0.73	0.72	0.74
Glu	0.59*	0.64*	0.65*	0.68*	0.68*
Phe	0.64	0.70	0.69*	0.67*	0.69*
Gly	0.80*	0.84*	0.84*	0.83*	0.85*
His	0.65	0.69	0.72	0.72	0.74
Ile	0.62*	0.67	0.73	0.63*	0.71
Lys	0.53*	0.62*	0.66*	0.65*	0.66*
Leu	0.67	0.69	0.68*	0.64*	0.68*
Met	0.65	0.72	0.66*	0.72	0.70
Asn	0.59*	0.65*	0.76	0.74	0.77
Pro	0.68	0.76*	0.74	0.76	0.76
Gln	0.62	0.70	0.71	0.71	0.74
Arg	0.69	0.65*	0.76	0.76*	0.76
Ser	0.61*	0.68	0.73	0.70	0.74
Thr	0.64	0.72	0.76	0.73	0.76
Val	0.62*	0.67*	0.74	0.67*	0.73
Trp	0.78*	0.80*	0.80	0.79	0.80
Tyr	0.67	0.68	0.72	0.69	0.72

However, while this task represents an important part of the use of these tools[7,34], it is not necessarily directly related to the task of estimating evolutionary constraint. It is also undeniably true that BARE cannot detect conserved noncoding elements, which is one common use of tools like phyloP and GERP++. In general, we do not consider BARE to be a strict improvement on these tools, but rather an additional resource to be used together with them.

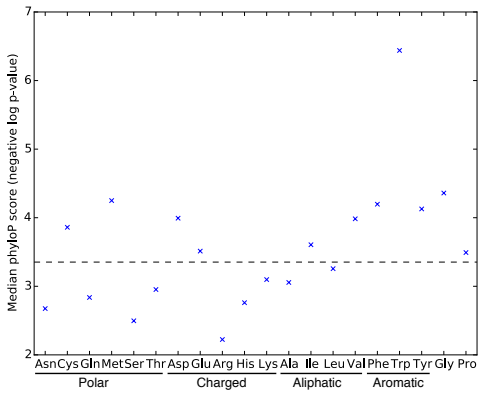
We have precomputed the BARE score for every protein-coding position in the human genome, as indexed by the UCSC genome browser database (`http://genome.ucsc.edu/`). These values are available at `http://genetics.bwh.harvard.edu/bare`, in the form of a genome browser track that can be displayed on the UCSC genome browser, a SQLite3 database, and a tab-delimited text file. We hope that making this resource available will provide the community with a new source of information about evolutionary constraint and the selective landscape experienced by protein-coding sequences.

In addition to these precomputed scores, we have released the source code to compile and run BARE as a standalone tool. The program takes as an argument a phylogenetic tree, an alignment of amino acid sequences, and a position in amino acid coordinates. We have exclusively used the UCSC MultiZ whole-genome orthologous alignments of 100 vertebrate species, along with its corresponding species tree; however, in principle any custom alignment or tree is accepted. Additionally, since it is possible to use a different tree for each position, BARE is in principle compatible with a wide variety of different tree models, including models involving widespread incomplete lineage sorting and horizontal gene transfer[43,202], as well as relaxed and local molecular clocks[19,46,174]. The trees are accepted in the standard Newick format, so a wide variety of external programs can be used to create these input trees. Similarly, the raw rate estimate produced by BARE can also be further processed to generate a variety of other scores that measure constraint, such as phyloP's constraint p-value or GERP++'s "rejected substitutions" score. In the case of GERP++ in particular, the BARE score can be used as a drop-in replacement for the native "neutral rate" score. This makes BARE a very

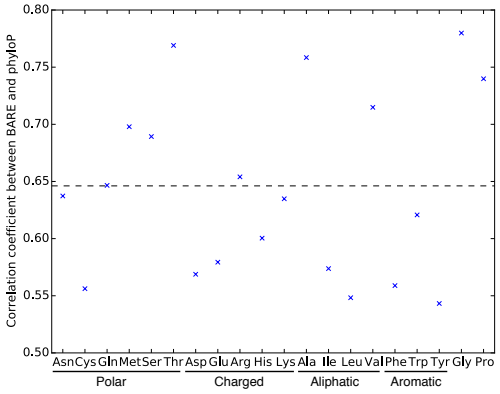
versatile tool with a wide variety of applications to computational genetics research.



(a)



(b)



(c)

Figure 3.2: Comparison of conservation scores for whole-exome datasets of BARE and phyloP scores (phyloP is used as a representative of nucleotide-level tree-based scores). Panels A and B show median scores from BARE (a) and phyloP (b). Higher values represent greater levels of conservation. Panel C shows correlation (Spearman rho) between BARE and phyloP. Dashed black line shows values for the complete unpartitioned dataset.

4

Conclusion

The field of variant effect prediction is in a somewhat confusing state. The general perception is that existing tools are immature and not to be trusted[71,153,176,205], leading to the release of new tools year after year[25,34,101,108,144]. Each new tool claims significant improvements over its predecessors, but independent analyses indicate that all methods have roughly comparable performance, with typical accuracy hovering around 80%[72,183]. In fact, despite the appearance of immaturity, these tools are extremely mature and rest on well-developed methodologies. The easy ways to improve perfor-

mance by incorporating new scores and new machine learning techniques have generally already been found; the remaining improvements are difficult.

In this dissertation, I have presented three of my attempts at these difficult improvements. In the first chapter, I improved the method's applicability to a specific clinical problem. In the second, I addressed the method's handling of epistasis. In the third, I incorporated a parametric model that explicitly models the tree structure of evolution. In all three cases, the results are not magical upgrades that dramatically improve performance. In chapter 1, the performance improvement was fairly dramatic, but this improvement required application of significant biological and biochemical insights. The most effective new score, the structure pair score, is actually impossible to compute for the vast majority of genes; the next most effective, the coiled-coil score, has no detectable predictive power when applied to the genome-wide HumVar dataset. The other major contributors to accuracy—manual curation of alignments and inclusion of the "no call" classification—are unique to the specific application and are not necessarily feasible to apply to a general-purpose or genome-wide predictor. The results from chapter 2 are arguably even more discouraging: my best attempt at predicting CPDs is utterly incapable of making a prediction without the aid of an *in vivo* model. Only chapter 3 has any kind of wide applicability to automated methods, and its performance improvement, though significant, is extremely modest. All in all, none of these advances are likely to be incorporated into the next version of PolyPhen. Instead, what makes the most difference in the performance of prediction methods are the continuous updating of databases of sequences and variants, not any theoretically interesting advance.

This should not necessarily be discouraging. Advances of the magnitude reported in chapter 1 may require specific biological knowledge and consulation with medical experts, but there is a great deal of such knowledge and expertise in the field. Furthermore, with the development of text mining algorithms and ontologies of biological models and phenotypes[107,121], this knowledge becomes easier and easier to access. Likewise, the method reported in chapter 2 requires *in vivo* experimenta-

tion to complement computational predictions, but these experiments continue to become faster and cheaper. Prioritizing variants to the point where the experiment can be followed up by *in vivo* methods is becoming a less and less onerous task. This kind of biological knowledge and experimental validation becomes more and more important as we move towards predicting more specific phenotypes and more complex genetic architectures, both of which are especially vital to the clinical use of these predictors. Additionally, as the amount of available data continues to grow, even complex and biologically specific problems like these become more and more tractable to automated methods.

In general, in order for these prediction methods to remain useful into the era of widespread clinical and population-level sequencing, they must become platforms for a wide range of more specific biological analyses, rather than attempting to be better at the general task of predicting the effects of variants. We should be skeptical of advances that claim to instantly improve performance by incorporating more data or new algorithms, and instead look for ways of extending and refining these already well-developed methods.

A

Supplementary Material to Chapter 2

The bulk of chapter 2 is a manuscript currently in press at *Nature*[99]. The material in this appendix will be available as an online supplement to that manuscript once it is published.

Supplementary Note: Additional characterization of CPDs

To characterize further our detected CPDs, we performed enrichment analysis on functional annotations and modes of inheritance. We performed Gene Ontology (GO) term enrichment analy-

sis using the AmiGO term enrichment tool (http://amigo1.geneontology.org/cgi-bin/amigo/term_enrichment), using the default thresholds of maximum p-value 0.01 (before multiple test correction) and minimum number of gene products 2. No significant term enrichments were found. We also performed annotation cluster enrichment analysis using the Database for Annotation, Visualization and Integrated Discovery (DAVID) functional annotation clustering tool[90]. This analysis showed modest enrichment for terms representing secreted extracellular compounds, integral membrane proteins, ion channels, and vacuoles. Other enrichments span a broad range of biological systems and processes, including blood clotting, blood lipid regulation, perception and cognition, hormone response, and embryonic development (Table S3).

Next, we performed enrichment analysis on modes of inheritance. Modes of inheritance were downloaded from Orphanet (http://www.orpha.net/), which labels diseases with inheritance categories including Autosomal Dominant, Autosomal Recessive, XLinked, Sporadic, and Multigenic/Multifactorial. In both the ClinVar and HumVar datasets, enrichment analysis found small but highly significant depletion for Autosomal Dominant, Autosomal Recessive, and X-Linked modes of inheritance, and significant enrichment for Multigenic/Multifactorial inheritance (Table S4). This is in keeping with our expectations, since diseases with multigenic and multifactorial modes of inheritance inherently involve multiple sites and thus should be more readily subject to compensation.

Finally, we searched dbSNP for candidate CPDs whose presence in the alignment may be explained by the variant being polymorphic in the other species. We found a total of 10 candidate CPDs that are present in dbSNP for another species (Table S5). Out of these, all but one were present in multiple species other than the one with a known polymorphism, suggesting that in general there are very few cases where a variant is found in the alignment solely because it is polymorphic. However, we cannot fully address the intersection of polymorphism and CPDs without more polymorphism data from a wider range of species.

These analyses complement previous studies that have examined the structural and chemical properties of CPDs, finding that CPDs are likely to be structurally destabilizing rather than impairing a specific biochemical function, and that they are likely to have milder effects on folding stability than other pathogenic SNVs[8,55].

Supplementary Note: CPD predictor usage

Based on our experience with the identification of a novel human disease gene containing a CPD, we suggest a Bayesian approach to CPD identification, supported by our probabilistic predictor. The *BTG2* V141M allele, identified as pathogenic in vivo but predicted as benign by multiple computational tools, should be considered initially an ambiguous case due to conflicting evidence. Our CPD predictor reports that it is unlikely but possible that this variant is a CPD, with probability 1.6%. This value represents the Bayesian posterior based solely on the alignment. However, by presenting human genetic evidence, demonstrating experimentally that the V141M allele is functional, and locating the compensating alleles that account for its presence in almost all orthologous sequences, we are able to plausibly claim that this variant falls in the 1.6% of variants seen in similar alignments that are CPDs, and dismiss the computational prediction as a false negative. If, on the other hand, the CPD predictor had reported a probability well below 1%, we might remain unconvinced even in the face of our genetic and functional evidence; meanwhile, if it had reported a probability well above 5%, the functional and genetic evidence might have been sufficient without identifying the compensating alleles. In our dataset, approximately 1,800 benign variants and 60 pathogenic variants are assigned a probability below 1%, while approximately 5,600 benign variants and 1,800 pathogenic variants are assigned a probability above 5%. The web interface outputs three distinct messages that repeat these recommendations.

Predicting specific compensating sites, though also desirable, is not feasible based on our current

knowledge. There are several theoretical models to explain the biochemical basis of compensatory events. These include reconstitution of destabilized tertiary structure, restoration of protein stability, and improvement of protein-protein interaction capabilities within a complex[8,47]. However, we do not know the biochemical basis of the compensatory events discovered here; for each of *BBS4*, *RPGRIP1L*, and *BTG2*, there was little *a priori* evidence for any of these interactions. The validated interactions can be long-range in terms of primary sequence, with one spanning more than 700 residues (Figure S3); none of the three proteins tested have known 3D structures; and none of these interactions suggest obvious biochemical mechanisms for rescue, such as balancing, electrostatic charge, or replacement of phosphorylation sites. It would be challenging for any computational method to account for these interactions and to make the correct prediction. Until more validated examples are collected, and/or until we have more biochemical information about those examples we have collected, predicting specific interactions in a principled way is not feasible. Instead, we have made available the method we used to design the experiments we reported here. This method simply treats any substitution that co-occurs with the candidate CPD as a candidate compensation, prioritizing sites that are substituted in multiple different species. These candidate lists may require fairly high-throughput experimental systems to test them. A more principled approach would not have this limitation.

Both the CPD prediction tool and the candidate compensation tool are available online at `http://genetics.bwh.harvard.edu/cpd/`. These tools should make it easier to interpret the output of computational tools like PolyPhen and SIFT, and to design experiments like those we report here. We expect future studies to develop more accurate predictors, possibly incorporating known functional and structural features of CPDs[8,55] and/or attempting to predict compensation sites in a principled way.

Supplementary Note: Manual evaluation of false-positive pathogenic alleles

To evaluate the false-positive annotation of variants as pathogenic in the unfiltered HumVar dataset, we selected 100 alleles randomly from the HumVar pathogenic alleles list, and another 100 random alleles from the HumVar compensated alleles list (a subset of the HumVar pathogenic alleles that are also found in other species). We inspected the evidence supporting the alleles manually, discarding any that were obvious false positives based on the following criteria:

a. *in vitro* or *in vivo* functional studies showed benign or only minor effects not significantly different from wild type;

b. no evidence of pathogenicity;

c. common polymorphisms (present in homozygosity in healthy controls, and/or minor allele frequency greater than 5%); or

d. incorrect annotation.

We found that 5/100 alleles in the HumVar pathogenic list and 6/100 alleles in the HumVar compensated list were annotated falsely as pathogenic.

Table A.1: Number of variants from all datasets shared between alignment strategies.

	MultiZ Unfiltered	MultiZ Mammals	MultiZ only EPO sequences	EPO	MultiZ High Quality	MultiZ >1 Species
MultiZ Unfiltered	16984	13766	11561	10135	13973	13174
MultiZ Mammals	13766	13766	11561	9698	11220	11681
MultiZ only EPO sequences	11561	11561	11561	9096	9360	10519
EPO	10135	9698	9096	11236	8339	9072
MultiZ High Quality	13973	11220	9360	8339	13973	10786
MultiZ>1 Species	13174	11681	10519	9072	10786	13174

Table A.2: Pathogenic human alleles present in multiple species.

(a) Present in both ClinVar and HumVar $\geq$ 2 species

Gene	Allele	Human Disease	Species	EVS	Inheritance	Onset	Reference	Comment
SCN5A	Val1951Met	Atrial fibrillation	59	absent	dominant	adolescence	40	
	Arg680His	Sudden unexplained death (SUD), Sudden infant death syndrome (SIDS)	43	absent	dominant	young adult	29, 189	
	Ala997Ser	Long QT Syndrome	10	absent	unknown	infancy	1	
ATRX	Leu409Ser	X-linked mental retardation	24	absent	X-linked	childhood	196	
AQP2	Leu22Val	Congenital nephrogenic diabetes insipidus (NDI)	17	absent	recessive	congenital	21	
	Ser216Pro	Congenital nephrogenic diabetes insipidus	2	absent	recessive	congenital	44, 45, 186	
ZNF81	Ser179Asn	X-linked mental retardation	9	absent	X-linked	childhood	109	
CFTR	Ile1234Val	Cystic fibrosis	8	absent	recessive	childhood	137	
TNNI3	Asn185Lys	Dilated cardiomyopathy	8	absent	dominant	childhood	23	
AMHR2	Arg406Gln	Persistent Mullerian duct syndrome	6	absent	recessive	congenital	12, 132	
OTC	Thr125Met	Ornithine carbamoyltransferase deficiency	5	TT=0/TC=1/CC=4058/C=2443	X-linked	congenital	66, 173	*cis*-compensation with 135T
TRPV4	Arg269His	Charcot-Marie-Tooth disease	5	absent	dominant	childhood	117	
FGFR3	Asp513Asn	Lacrimo-auriculo-dento-digital (LADD) syndrome	3	absent	dominant	congenital	160	
FIG4	Ile41Thr	Charcot-Marie-Tooth disorder, type 4J	2	CC=0/CT=15/TT=6487	recessive	childhood	32, 119	
PDE8B	His305Pro	Bilateral adrenocortical hyperplasia/Cushing syndrome	2	absent	dominant	childhood	85, 86	
AIRE	Leu29Pro	autoimmune polyendocrinopathy (APE)-candidiasis (C)-ectodermal dystrophy (ED), APECED	2	absent	recessive	infancy	78, 110	

Table A.2: (continued)

(b) Present in both ClinVar and HumVar 1 species

Gene	Allele	Human Disease	Species	EVS	Inheritance	Onset	Reference	Comment
LHX4	Arg84Cys	Growth hormone deficiency	1	absent	recessive	childhood	151	
PROP1	Arg99Gln	Combined pituitary hormone deficiency (CPHD)	1	absent	recessive	adolescence	187	
POF1B	Arg329Gln	Premature ovarian failure (POF)	1	TT=0/TC=35/T=8/CC=4025/C=2435	X-linked	adolescence	116, 149	
GCM2	Gly63Ser	Familial isolated hypoparathyroidism	1	TT=0/TC=1/CC=6502	recessive	childhood	127, 182	
NOG	Tyr222Cys	Proximal symphalangism and multiple synostoses syndrome	1	absent	dominant	congenital	50, 70	
SI	Gln117Arg	Congenital sucrase-isomaltase deficiency (CSID)	1	absent	unknown	congenital	169	

Table A.2: (continued)

(c) Present in ClinVar $\geq$ 3 species

Gene	Allele	Human Disease	Species	EVS	Inheritance	Onset	Reference	Comment
PRSS1	Asn29Thr	Hereditary chronic pancreatitis	68	absent	dominant	childhood	152, 166	
	Arg122His	Hereditary chronic pancreatitis	5	absent	dominant	childhood	28, 177, 194, 197	
CD96	Thr280Met	C Syndrome/Opitz Trigonocephaly	27	absent	de novo	infancy	102	
KCNA5	Ala576Val	Atrial fibrillation	25	absent	dominant	adolescence	200	
LTBP2	Val1177Met	Weill-Marchesani syndrome	9	absent	recessive	congenital	77	
NR5A1	Gly212Ser	Spermatogenic failure	9	TT=0/TC=1/CC=6280	de novo	adolescence	9	
KISS1R	Arg386Pro	Central precocious puberty	6	absent	de novo	adolescence	14, 178	
HSD17B3	Ala203Val	Testosterone 17-beta-dehydrogenase deficiency/Male pseudohermaphroditism	6	absent	recessive	congenital	61	
CYP21A2	Pro30Leu	21-hydroxylase deficiency	6	absent	recessive	congenital	131, 147, 168, 185	
	Pro105Leu	Congenital adrenal hyperplasia (CAH)	4	absent	recessive	congenital	146	
ABCB6	Ala57Thr	Microphthalmia, isolated, with coloboma	5	absent	dominant	congenital	191	
ELANE	Val98Leu	Severe congenital and cyclic neutropenia	5	absent	recessive	congenital	65, 157, 163	called V69L
TBC1D24	Phe251Leu	Familial infantile myoclonic epilepsy	3	absent	recessive	childhood	35	
NEFL	Pro22Ser	Autosomal dominant Charcot-Marie-Tooth disease	3	absent	dominant	childhood	54, 63, 164	

Table A.2: (continued)

(d) Present in HumVar $\geq$ 3 species

Gene	Allele	Human Disease	Species	EVS	Inheritance	Onset	Reference	Comment
ATP7B	Met1169Val	Wilson disease	74	absent	recessive	childhood	88	
GH1	Ser134Arg	Dwarfism/short stature	59	absent	dominant	childhood	134	called S108R
GLA	Arg356Gln	Fabry disease	11	absent	X-linked	childhood	94, 162	
COLQ	Arg410Gln	End-plate acetylcholinesterase deficiency	8	absent	recessive	congenital	148	
SUMF1	Leu20Phe	Multiple sulfatase deficiency	7	absent	unknown	childhood	37	
MUTYH	Arg182Trp	Familial adenomatous polyposis	5	absent	recessive	childhood	136	
NODAL	Glu203Lys	Autosomal visceral heterotaxy/cardiovascular malformations	5	absent	sporadic	congenital	135	
KAL1	Phe517Leu	Hypogonadotropic hypogonadism	3	absent	X-linked	childhood	24, 64, 89	
NAGLU	Ala246Pro	Mucopolysaccharidosis 3B	3	absent	recessive	childhood	10	
FOXE3	Gly49Ala	Anophthalmia and microphthalmia	3	absent	dominant	congenital	96	
F8	His1066Tyr	Hemophilia A	3	absent	X-linked	congenital	76, 150	called H1047Y

(e) Present in ClinVar 1–2 species

Gene	Allele	Human Disease	Species	EVS	Inheritance	Onset	Reference	Comment
HSD11B2	Asp223Asn	Apparent Mineralocorticoid Excess (AME)	2	absent	recessive	congenital	26	
CD79B	Gly138Ser	Immunodeficiency	2	absent	recessive	childhood	51	called G137S
TRPA1	Asn855Ser	Autosomal-dominant pain syndrome (FEPS)	2	absent	dominant	congenital	114	

(f) Present in HumVar 1–2 species

Gene	Allele	Human Disease	Species	EVS	Inheritance	Onset	Reference	Comment
ITGA2B	Pro943Leu	Type II Glanzmann thrombasthenia	2	absent	recessive	childhood	97	called P912L
SLC2A9	Arg171Cys	Renal hypouricemia type 2	2	absent	recessive	childhood	48	
ARSA	Ala18Asp	Metachromatic leukodystrophy (infantile form)	2	absent	recessive	infancy	73	
SLC7A9	Ala70Val	non-Type I cystinuria	2	absent	recessive	childhood	58	
ARSB	Cys192Arg	Type VI mucopolysaccharidosis	1	absent	recessive	childhood	95, 199	

Table A.3: Top 50 annotation clusters identified by DAVID analysis

Cluster number	Representative annotation	Enrichment score
1	secreted	15.3
2	blood coagulation	13.1
3	visual perception	12.8
4	blood circulation	11.8
5	response to hormone stimulus	10.2
6	chordate embryonic development	10.2
7	lytic vacuole	9.21
8	protein homodimerization activity	9.05
9	ion homeostasis	6.00
10	response to nutrient levels	8.61
11	muscle contraction	7.83
12	steroid biosynthetic process	7.54
13	kidney development	6.88
14	membrane fraction	6.87
15	limb development	6.11
16	intrinsic to plasma membrane	5.92
17	heart development	5.87
18	cellular amino acid derivative metabolic process	5.87
19	neurodegeneration	5.47
20	eye development	5.31
21	sex differentiation	5.29
22	EGF calcium-binding	5.18

Table A.3 (continued)

Cluster number	Representative annotation	Enrichment score
23	oxidation reduction	5.14
24	vesicle	5.00
25	neural tube development	4.99
26	forebrain development	4.98
27	vasculature development	4.89
28	structural constituent of cytoskeleton	4.85
29	glycation	4.83
30	pigment biosynthetic process	4.58
31	acute inflammatory response	4.57
32	ion transport	4.25
33	respiratory system development	4.15
34	response to carbohydrate stimulus	4.12
35	platelet alpha granule	4.06
36	cell adhesion	3.90
37	collagen	3.88
38	vacuole organization	3.78
39	regulation of protein modification process	3.76
40	cilium	3.72
41	metalloprotein	3.71
42	bone development	3.61
43	transmission of nerve impulse	3.58
44	regulation of neurological system process	3.54

Table A.3 (continued)

Cluster number	Representative annotation	Enrichment score
45	steroid biosynthesis	3.48
46	endoplasmic reticulum	3.46
47	folate biosynthesis	3.41
48	regulation of blood vessel size	3.41
49	cholesterol metabolic process	3.41
50	sulfatase	3.38

Table A.4: Mode of inheritance enrichment analysis.

	ClinVar Pathogenic	ClinVar CPDs	Depletion p-value	HumVar Pathogenic	HumVar CPDs	Depletion p-value
Autosomal Dominant	2713	252	0.037	4359	432	1.07×10^{-6}
Autosomal Recessive	2773	240	0.00068	5945	633	0.00011
X-Linked	586	35	0.00015	2463	138	1.35×10^{-29}
Sporadic	96	13	0.895	227	16	0.011
Multigenic/Multifactorial	40	9	0.9945	43	11	0.9964

Table A.5: Candidate CPD sites with known polymorphisms in other species.

Species	rsID	Human Variant	Species in alignment
Mouse	rs8249104	DMBT1:p.Asn546Ser	38
Mouse	rs2290114447	HFE:p.Gln127His	8
Mouse	rs258538416	ABCB4:p.Ile764Leu	9
Mouse	rs222883690	USH2A:p.Arg2354His	32
Pig	rs325593396	MEFV:p.Phe479Leu	57
Pig	rs336797577	PKHD1:p.Tyr2661His	13
Sheep	rs406745685	EYS:p.Leu2189Pro	4
Sheep	rs160536391	JAG1:p.Ser913Arg	1
Dog	rs22977833	HBB:p.Val12Ile	37
Chicken	rs314119972	TNFRSF13B:p.Ala181Gly	4

Table A.6: Significance of differences between pathogenic and benign distributions for HumVar and ClinVar and various alignment strategies (Kolmogorov-Smirnof 2-sample test).

	HumVar	ClinVar
MultiZ Unfiltered	5.5×10^{-173}	2.0×10^{-42}
MultiZ Mammals	1.6×10^{-68}	7.0×10^{-17}
EPO	3.1×10^{-124}	3.0×10^{-19}
MultiZ Quality Filtered	2.0×10^{-153}	3.0×10^{-14}
MultiZ >1 Sequences	3.3×10^{-82}	2.4×10^{-17}

Table A.7: Significance of differences between pathogenic variant distributions for MultiZ alignment (mammals only) (Kolmogorov-Smirnof 2-sample test).

	HumVar	ClinVar	HumVar + ClinVar	HumVar + ClinVar excluding ESP
HumVar		0.46	0.77	0.52
ClinVar	0.46		0.42	0.30
HumVar + ClinVar	0.77	0.42		1.00
HumVar + ClinVar excluding ESP	0.52	0.30	1.00	

Table A.8: Fitted parameters for models. ∗ indicates value fixed by the model, not fitted independently. Models are described in the methods section of the main text. Values are maximum likelihood estimates; error ranges are standard errors of the likelihood distribution. Intersected datasets (HumVar+ClinVar, HumVar+ClinVar+ESP) use HumVar annotations for neutral variants. Time is measured in sequence distance (1-sequence identity).

model	variant	alignment	mean time to fix neutral variants	mean time to fix CPDs	mean time to fix compensations	number of compensatory sites
Model 1	HumVar	MultiZ	0.17 ± 0.001	0.11 ± 0.003	0.11 ± 0.003∗	1.4 ± 0.07
		mammals	0.14 ± 0.001	0.075 ± 0.003	0.075 ± 0.003∗	1.5 ± 0.08
		EPO	0.14 ± 0.001	0.14 ± 0.006	0.14 ± 0.006∗	1.7 ± 0.06
		quality	0.16 ± 0.002	0.11 ± 0.004	0.11 ± 0.004∗	1.4 ± 0.07
		# hits	0.14 ± 0.001	0.092 ± 0.004	0.092 ± 0.004∗	1.4 ± 0.09
	ClinVar	MultiZ	0.16 ± 0.005	0.12 ± 0.006	0.12 ± 0.006∗	1.2 ± 0.1
		mammals	0.12 ± 0.005	0.083 ± 0.006	0.083 ± 0.006∗	1.2 ± 0.1
		EPO	0.13 ± 0.006	0.17 ± 0.01	0.17 ± 0.01∗	1.4 ± 0.08
		quality	0.14 ± 0.006	0.12 ± 0.007	0.12 ± 0.007∗	1.2 ± 0.1
		# hits	0.13 ± 0.005	0.10 ± 0.007	0.10 ± 0.007∗	1.1 ± 0.1

Table A.8 (continued)

model	variant	alignment	mean time to fix neutral variants	mean time to fix CPDs	mean time to fix compensations	number of compensatory sites
	HumVar / ClinVar	MultiZ	0.17 ± 0.001	0.11 ± 0.009	0.11±0.009*	1.5 ± 0.2
		mammals	0.14 ± 0.001	0.083 ± 0.009	0.083 ± 0.009*	1.4 ± 0.2
		EPO	0.14 ± 0.001	0.18 ± 0.05	0.18 ± 0.05*	0.63 ± 0.2
		quality	0.16 ± 0.002	0.11 ± 0.009	0.11±0.009*	1.5 ± 0.2
		# hits	0.14 ± 0.001	0.10 ± 0.01	0.10 ± 0.01*	1.2 ± 0.3
	HumVar / ClinVar excluding ESP	MultiZ	0.17 ± 0.001	0.11 ± 0.01	0.11 ± 0.01*	1.6 ± 0.2
		mammals	0.14 ± 0.001	0.089 ± 0.01	0.089 ± 0.01*	1.3 ± 0.3
		EPO	0.14 ± 0.001	0.18 ± 0.06	0.18 ± 0.06*	0.71 ± 0.2
		quality	0.16 ± 0.002	0.11 ± 0.01	0.11 ± 0.01*	1.7 ± 0.3
		# hits	0.14 ± 0.001	0.10 ± 0.02	0.10 ± 0.02*	1.3 ± 0.3
Model 2	HumVar	MultiZ	0.17 ± 0.001	0.11 ± 0.008	0.11 ± 0.01	1.4 ± 0.07
		mammals	0.14 ± 0.001	0.075 ± 0.01	0.076 ± 0.009	1.5 ± 0.09
		EPO	0.14 ± 0.001	0.15 ± 0.02	0.14 ± 0.02	0.68 ± 0.08
		quality	0.16 ± 0.002	0.11 ± 0.01	0.11 ± 0.01	1.4 ± 0.07
		# hits	0.14 ± 0.001	0.09 ± 0.01	0.092 ± 0.009	1.4 ± 0.09

Table A.8 (continued)

model	variant	alignment	mean time to fix neutral variants	mean time to fix CPDs	mean time to fix compensations	number of compensatory sites
	ClinVar	MultiZ	0.16 ± 0.005	0.12 ± 0.02	0.13 ± 0.02	1.2 ± 0.1
		mammals	0.12 ± 0.005	0.082 ± 0.03	0.083 ± 0.02	1.3 ± 0.1
		EPO	0.13 ± 0.006	0.20 ± 0.03	0.15 ± 0.06	0.19 ± 0.2
		quality	0.14 ± 0.006	0.12 ± 0.02	0.12 ± 0.01	1.2 ± 0.1
		# hits	0.13 ± 0.005	0.10 ± 0.02	0.10 ± 0.02	1.1 ± 0.1
	HumVar / ClinVar	MultiZ	0.17 ± 0.001	0.11 ± 0.03	0.11 ± 0.02	1.5 ± 0.2
		mammals	0.14 ± 0.001	0.082 ± 0.05	0.083 ± 0.03	1.4 ± 0.3
		EPO	0.14 ± 0.001	0.18 ± 0.05	0.15 ± 0.07	0.55 ± 0.2
		quality	0.16 ± 0.002	0.11 ± 0.02	0.11 ± 0.03	1.5 ± 0.2
		# hits	0.14 ± 0.001	0.099 ± 0.04	0.099 ± 0.04	1.2 ± 0.3
	HumVar / ClinVar excluding ESP	MultiZ	0.17 ± 0.001	0.10 ± 0.03	0.11 ± 0.02	1.6 ± 0.2
		mammals	0.14 ± 0.001	0.088 ± 0.05	0.089 ± 0.04	1.3 ± 0.3
		EPO	0.14 ± 0.001	0.18 ± 0.08	0.14 ± 0.1	0.66 ± 0.3
		quality	0.16 ± 0.002	0.10 ± 0.03	0.11 ± 0.02	1.7 ± 0.3
		# hits	0.14 ± 0.001	0.10 ± 0.05	0.10 ± 004	1.3 ± 0.3
Model 3	HumVar	MultiZ	0.17 ± 0.001	0.17 ± 0.001*	0.078 ± 0.006	1.4 ± 0.09
		mammals	0.14 ± 0.001	0.14 ± 0.001*	0.034 ± 0.005	1.9 ± 0.2

Table A.8 (continued)

model	variant	alignment	mean time to fix neutral variants	mean time to fix CPDs	mean time to fix compensations	number of compensatory sites
		EPO	0.14 ± 0.001	0.14±0.001*	0.15 ± 0.01	0.71 ± 0.06
		quality	0.14 ± 0.002	0.14 ± 0.002*	0.077 ± 0.006	1.5 ± 0.1
		# hits	0.13 ± 0.005	0.13±0.005*	0.059 ± 0.007	1.5 ± 0.1
	ClinVar	MultiZ	0.15 ± 0.005	0.15±0.005*	0.11 ± 0.01	1.2 ± 0.1
		mammals	0.12 ± 0.004	0.12 ± 0.004*	0.051 ± 0.01	1.4 ± 0.3
		EPO	0.13 ± 0.006	0.13 ± 0.006*	0.23 ± 0.04	0.41 ± 0.09
		quality	0.14 ± 0.006	0.14 ± 0.006*	0.11 ± 0.01	1.2 ± 0.1
		# hits	0.13 ± 0.005	0.13±0.005*	0.077 ± 0.02	1.1 ± 0.2
	HumVar / ClinVar	MultiZ	0.17 ± 0.001	0.17±0.001*	0.080 ± 0.02	1.4 ± 0.2
		mammals	0.14 ± 0.001	0.14±0.001*	0.037 ± 0.02	1.9 ± 0.6
		EPO	0.14 ± 0.001	0.14±0.001*	0.18 ± 0.05	0.63 ± 0.2
		quality	0.16 ± 0.002	0.16 ± 0.002*	0.083 ± 0.02	1.5 ± 0.2
		# hits	0.14 ± 0.001	0.14±0.001*	0.065 ± 0.02	1.3 ± 0.4

Table A.8 (continued)

model	variant	alignment	mean time to fix neutral variants	mean time to fix CPDs	mean time to fix compensations	number of compensatory sites
	HumVar / ClinVar excluding ESP	MultiZ	0.17 ± 0.001	0.17±0.001*	0.086 ± 0.02	1.5 ± 0.3
		mammals	0.14 ± 0.001	0.14±0.001*	0.049 ± 0.02	1.6 ± 0.6
		EPO	0.14 ± 0.001	0.14±0.001*	0.18 ± 0.06	0.71 ± 0.2
		quality	0.16 ± 0.002	0.16 ± 0.002*	0.086 ± 0.02	1.6 ± 0.3
		# hits	0.14 ± 0.001	0.14±0.001*	0.080 ± 0.03	1.3 ± 0.4

Table A.9: *BBS4* and *RPGRIP1L* candidate compensatory sites and *in vivo* assessment in zebrafish.

injection	normal	class 1	class 2	total	p-value vs MO	p-value vs WT rescue
UI Ctrl	338	11	2	351	<0.0001	<0.0001
bbs4 MO	87	267	78	432	—	<0.0001
bbs4 MO + WT RNA	133	39	4	176	<0.0001	—
bbs4 MO + N165H	52	91	38	181	0.027	<0.0001
bbs4 MO + N160H/N165H	32	44	20	96	0.0078	<0.0001
bbs4 MO + N160L/N165H	28	64	24	116	0.4166	<0.0001
bbs4 MO + N160R/N165H	26	88	14	128	0.9203	<0.0001
bbs4 MO + N160S/N165H	18	52	20	90	0.92	<0.0001
bbs4 MO + N163Q/N165H	32	44	16	92	0.0037	<0.0001
bbs4 MO + N165H/H366N	26	70	22	118	0.7518	<0.0001
bbs4 MO + N165H/H366R	84	32	0	116	<0.0001	0.639
bbs4 MO + N165H/H366S	32	90	14	136	0.4666	<0.0001
bbs4 MO + N165H/H366T	66	29	12	107	<0.0001	0.019
bbs4 MO + H366R	84	16	18	118	<0.0001	0.4839
bbs4 MO + H366T	74	14	12	100	<0.0001	0.8875
UI Ctrl	1058	30	3	1091	<0.0001	0.0003
rpgrip1l MO	280	207	76	563	—	<0.0001
rpgrip1l MO + WT RNA	134	11	3	148	<0.0001	—
rpgrip1l MO + R937L	32	29	4	65	0.9203	<0.0001
rpgrip1l MO + R94Q/R937L	94	60	6	160	0.0544	<0.0001
rpgrip1l MO + K124R/R937L	72	66	8	146	1	<0.0001
rpgrip1l MO + T128V/R937L	66	66	18	150	0.2117	<0.0001

Table A.9 (continued)

injection	normal	class 1	class 2	total	p-value vs MO	p-value vs WT rescue
rpgrip1l MO + T142I/R937L	13	15	23	51	0.0015	<0.0001
rpgrip1l MO + K179R/R937L	14	28	20	62	<0.0001	<0.0001
rpgrip1l MO + P189L/R937L	41	12	5	58	0.006	0.0008
rpgrip1l MO + H190Q/R937L	29	20	10	59	1	<0.0001
rpgrip1l MO + F193L/R937L	45	15	2	62	0.009	0.0017
rpgrip1l MO + A284V/R937L	31	22	13	66	0.7642	<0.0001
rpgrip1l MO + K335T/R937L	26	20	10	56	0.7401	<0.0001
rpgrip1l MO + D367N/R937L	25	22	7	54	0.729	<0.0001
rpgrip1l MO + V421I/R937L	21	17	7	45	0.8065	<0.0001
rpgrip1l MO + K443N/R937L	21	22	7	50	0.3681	<0.0001
rpgrip1l MO + L447F/R937L	29	17	5	51	0.4062	<0.0001
rpgrip1l MO + V481E/R937L	33	18	4	55	0.1897	<0.0001
rpgrip1l MO + R535Q/R937L	27	18	9	54	0.9203	<0.0001
rpgrip1l MO + H563Q/R937L	31	24	13	68	0.6033	<0.0001
rpgrip1l MO + V647I/R937L	35	25	14	74	0.7913	<0.0001
rpgrip1l MO + R649Q/R937L	25	17	10	52	0.9203	<0.0001
rpgrip1l MO + E689D/R937L	26	19	12	57	0.6468	<0.0001
rpgrip1l MO + I717V/R937L	24	14	10	48	0.9203	<0.0001
rpgrip1l MO + Q903K/R937L	17	24	8	49	0.0614	<0.0001
rpgrip1l MO + R937L/L958V	25	16	12	53	0.8231	<0.0001
rpgrip1l MO + R937L/R961T	42	7	8	57	0.0009	0.004
rpgrip1l MO + R937L/D986S	24	19	13	56	0.3994	<0.0001

Table A.9 (continued)

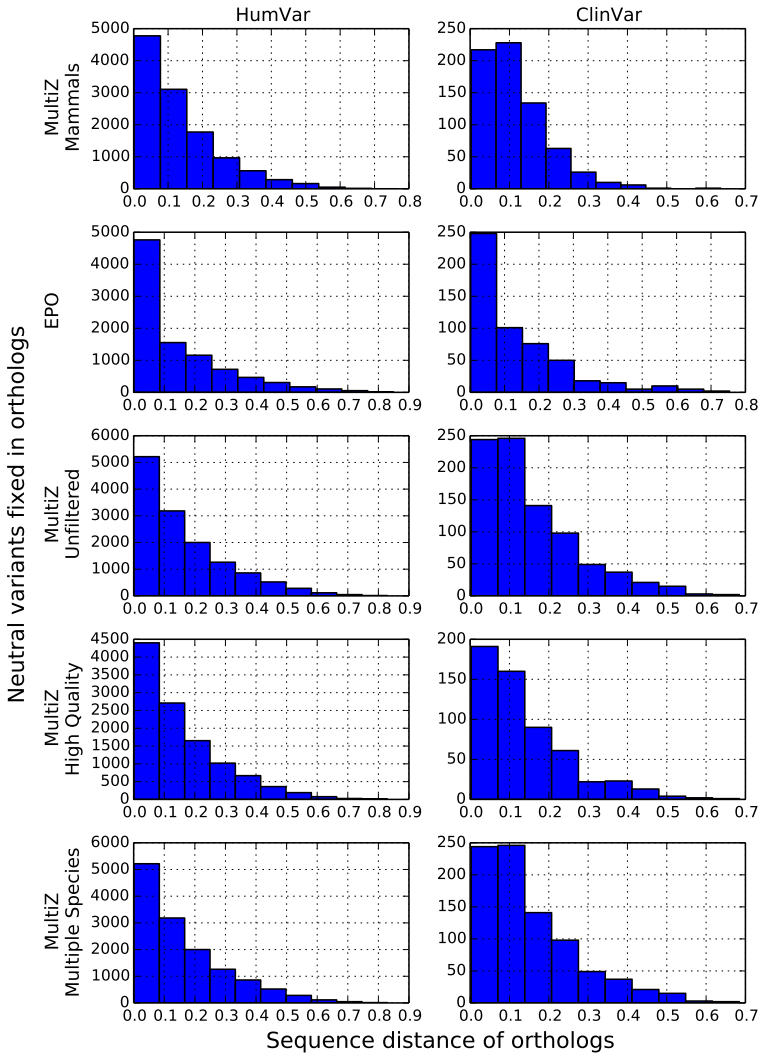
injection	normal	class 1	class 2	total	p-value vs MO	p-value vs WT rescue
rpgrip1l MO + R937L/P1016S	20	20	20	60	0.0226	<0.0001
rpgrip1l MO + R937L/K1020E	26	15	15	56	0.7401	<0.0001
rpgrip1l MO + R937L/I1092V	26	22	15	63	0.2542	<0.0001
rpgrip1l MO + R937L/Q1162R	19	24	12	55	0.0444	<0.0001
rpgrip1l MO + R937L/K1215Q	32	21	15	68	0.7773	<0.0001
rpgrip1l MO + R937L/V1257I	21	25	11	57	0.0859	<0.0001
rpgrip1l MO + R937L/L1272I	26	17	18	61	0.3566	<0.0001
rpgrip1l MO + P189L	68	18	6	92	<0.0001	0.0012
rpgrip1l MO + F193L	54	20	4	78	0.0019	0.0001
rpgrip1l MO + R961T	76	26	6	108	<0.0001	<0.0001

Table A.10: Head size measurement summary and p-values for *btg2*, *nos2a/b* MO-injected embryos. n/a, not applicable; MO, morpholino; UI, uninjected.

Injection	Number of Embryos	Distance from forebrain to hindbrain		p-value vs. UI control	p-value vs. *btg2* MO	p-value vs. *BTG2* WT rescue
		Mean distance (μm)	Standard deviation			
Control	79	623.06	27.16	n/a		
btg2 MO	73	566.51	36.44	<0.0001	n/a	
btg2 MO + *BTG2* WT mRNA	54	614.95	30.68	0.11	<0.0001	n/a
btg2 MO + BTG2 p.V141M mRNA	72	585.27	28.23	<0.001	0.0006	<0.0001
BTG2 WT mRNA	49	626.32	24.37	0.50	<0.0001	n/a
BTG2 p.V141M mRNA	72	613.60	26.62	0.30	<0.0001	n/a
nos2a MO	77	615.06	31.29	0.40	n/a	n/a
nos2b MO	53	608.46	30.85	0.15	n/a	n/a

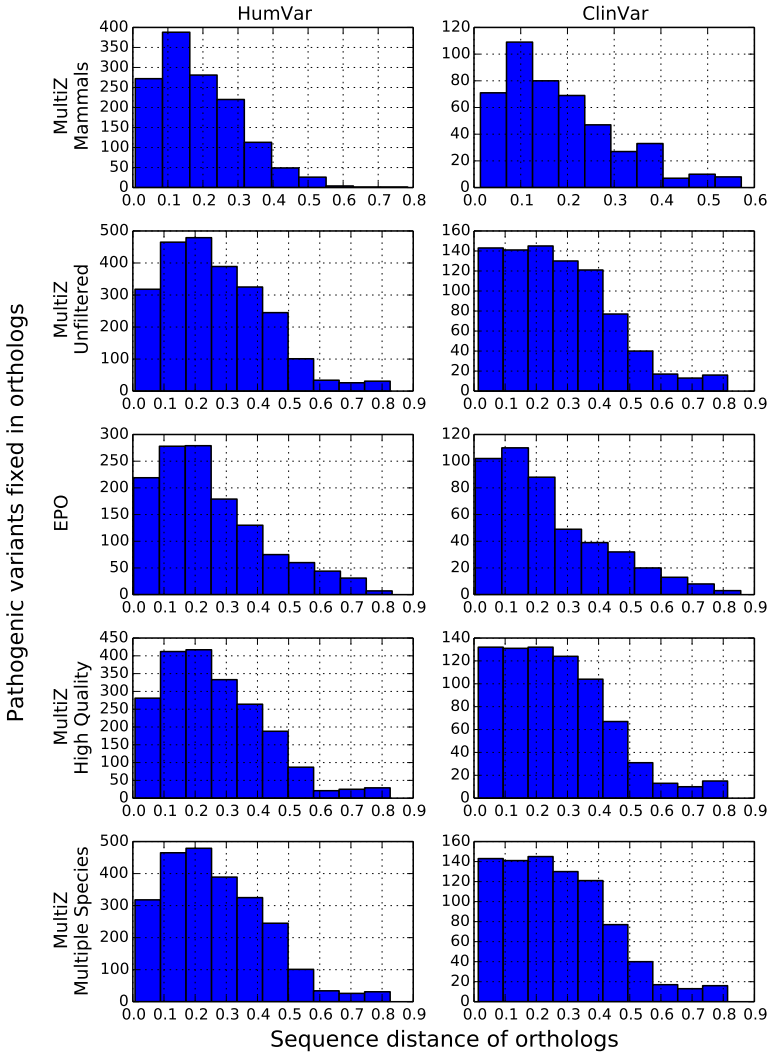
Table A.11: *BTG2* candidate compensatory sites and *in vivo* quantification of proliferating cells in the zebrafish head as determined by phospho-H3 immunostaining.

BTG2 Injection	Average cell count	StdDev	Number of embryos	p-value vs MO	p-value vs WT rescue
UI Ctrl	386.65	55.85	26	<0.0001	0.0104
btg2 MO	234.96	69.00	26	—	<0.0001
MO + WT RNA	357.27	61.73	22	<0.0001	—
MO + V141M	220.86	65.75	14	0.3759	0.0005
MO + V141M/G6R	206.44	58.38	16	0.2493	<0.0001
MO + V141M/G40R	247.15	57.55	20	0.1496	<0.0001
MO + V141M/R80K	378.05	71.87	22	<0.0001	0.2004
MO + V141M/Q94R	268.91	47.49	11	0.2249	0.0048
MO + V141M/S98R	250.92	87.87	13	0.3804	0.0044
MO + V141M/L128V	368.24	95.68	17	<0.0001	0.0969
MO + V141M/A130T	234.80	61.20	10	0.2677	0.0018
MO + V141M/C132Y	240.33	59.26	12	0.3023	0.0002
MO + V141M/L142M	227.20	29.91	10	0.1108	<0.0001
MO + R80K	326.20	60.63	10	0.0048	0.2091
MO + L128V	334.75	46.65	12	<0.0001	0.4383



(a) Distributions of neutral variants.

Figure A.1: Different alignment methodologies with HumVar and ClinVar produce qualitatively similar alignments. A–B) Distributions of missense variants annotated as neutral (A) or pathogenic (B) in the HumVar and ClinVar datasets, with each of the five alignment strategies described in the text (MultiZ unfiltered, MultiZ mammals-only, EPO, MultiZ with alignment quality filter, MultiZ with >1 sequence filter). All distributions are qualitatively similar. Compare to Figures 2C–D of the main text.



(b) Distributions of pathogenic variants.

Figure A.1: (continued)

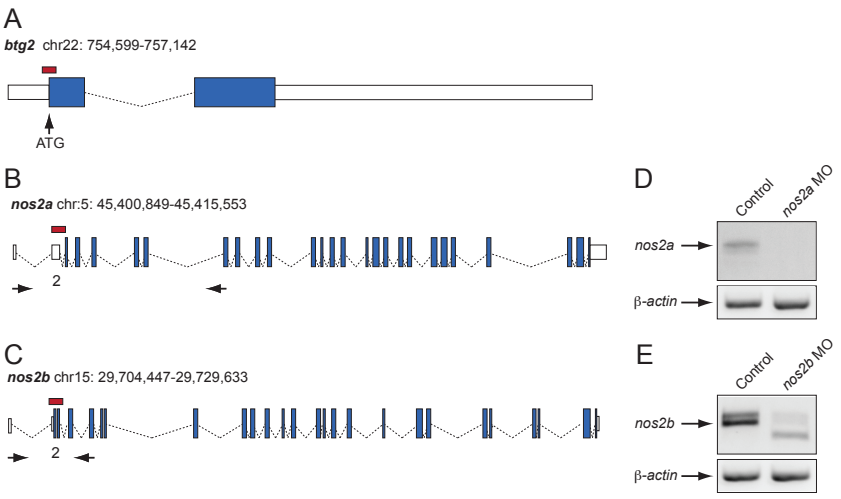


Figure A.2: Evaluation of btg2 and nos2a/b morpholinos (MO)s. A, B, C) Schematic of the D. rerio *btg2*, *nos2a* and *nos2b* loci. Blue boxes, exons; dashed lines, introns; white boxes, untranslated regions; red boxes, MOs; ATG indicates the translational start site; arrows, RT-PCR primers; number indicates the targeted exon. D, E) Agarose gel images of *nos2a/b* RT-PCR products.

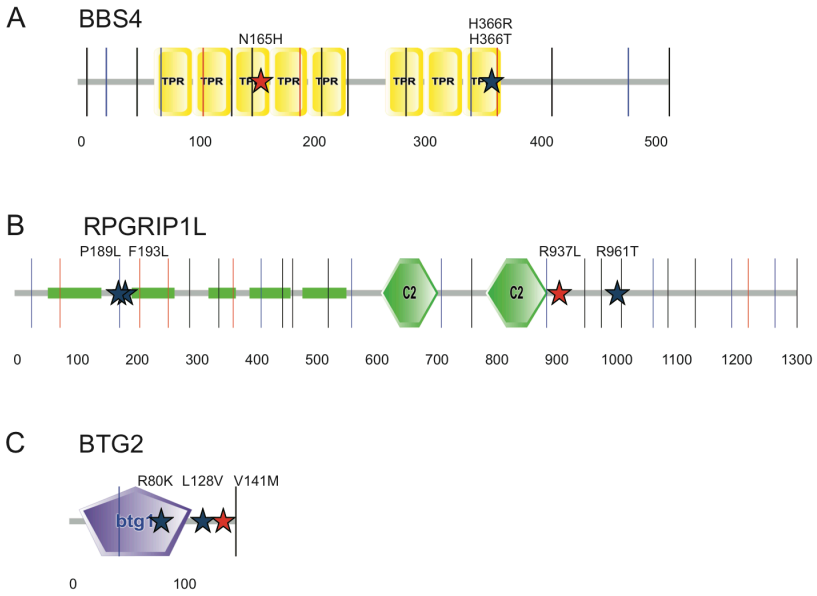


Figure A.3: Protein domain structure of functionally tested human disease genes. A) Schematic of *BBS4* (519 amino acids) is depicted with eight tetratricopeptide (TPR) domains (yellow); B) RPGRIP1L (1315 amino acids) has multiple coiled-coil domains (green rectangles) and two Protein kinase C conserved region 2 (C2) domains (green hexagons); and C) *BTG2* (158 amino acids) has one *BTG1* domain (purple pentagon). Disease-causing alleles are shown with red stars; complementing alleles are represented with blue stars; amino acid number scale in increments of 100 is shown below each schematic.

References

[1] Ackerman, M. J., Siu, B. L., Sturner, W. Q., Tester, D. J., Valdivia, C. R., Makielski, J. C., & Towbin, J. A. (2001). Postmortem molecular analysis of SCN5A defects in sudden infant death syndrome. *JAMA*, 286(18), 2264–2269.

[2] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods*, 7(4), 248–249.

[3] Ahola, V., Aittokallio, T., Vihinen, M., & Uusipaikka, E. (2008). Model-based prediction of sequence alignment quality. *Bioinformatics*, 24(19), 2165–2171.

[4] Alföldi, J. & Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. *Genome Res.*, 23(7), 1063–1068.

[5] Andersen, P. S., Havndrup, O., Bundgaard, H., Larsen, L. A., Vuust, J., Pedersen, A. K., Kjeldsen, K., & Christiansen, M. (2004). Genetic and phenotypic characterization of mutations in myosin-binding protein c (mybpc3) in 81 families with familial hypertrophic cardiomyopathy: total or partial haploinsufficiency. *Eur J Hum Genet*, 12(8), 673–677.

[6] Ashenberg, O., Gong, L. I., & Bloom, J. D. (2013). Mutational effects on stability are largely conserved during protein evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 110(52), 21071–21076.

[7] Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, 12(11), 745–755.

[8] Baresić, A., Hopcroft, L. E. M., Rogers, H. H., Hurst, J. M., & Martin, A. C. R. (2010). Compensated Pathogenic Deviations: Analysis of Structural Effects. *J. Mol. Biol.*, 396(1), 19–30.

[9] Bashamboo, A., Ferraz-de Souza, B., Lourenço, D., Lin, L., Sebire, N. J., Montjean, D., Bignon-Topalovic, J., Mandelbaum, J., Siffroi, J.-P., Christin-Maitre, S., Radhakrishna, U., Rouba, H., Ravel, C., Seeler, J., Achermann, J. C., & McElreavey, K. (2010). Human male infertility associated with mutations in NR5A1 encoding steroidogenic factor 1. *Am. J. Hum. Genet.*, 87(4), 505–512.

[10] Beesley, C. E., Jackson, M., Young, E. P., Vellodi, A., & Winchester, B. G. (2005). Molecular defects in Sanfilippo syndrome type B (mucopolysaccharidosis IIIB). *J. Inherit. Metab. Dis.*, 28(5), 759–767.

[11] Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675), 1321–1325.

[12] Belville, C., Maréchal, J.-D., Pennetier, S., Carmillo, P., Masgrau, L., Messika-Zeitoun, L., Galey, J., Machado, G., Treton, D., Gonzalès, J., Picard, J.-Y., Josso, N., Cate, R. L., & di Clemente, N. (2009). Natural mutations of the anti-Mullerian hormone type II receptor found in persistent Mullerian duct syndrome affect ligand binding, signal transduction and cellular transport. *Hum. Mol. Genet.*, 18(16), 3002–3013.

[13] Beunders, G., Voorhoeve, E., Golzio, C., Pardo, L. M., Rosenfeld, J. A., Talkowski, M. E., Simonic, I., Lionel, A. C., Vergult, S., Pyatt, R. E., van de Kamp, J., Nieuwint, A., Weiss, M. M., Rizzu, P., Verwer, L. E. N. I., van Spaendonk, R. M. L., Shen, Y., Wu, B.-l., Yu, T., Yu, Y., Chiang, C., Gusella, J. F., Lindgren, A. M., Morton, C. C., van Binsbergen, E., Bulk, S., van Rossem, E., Vanakker, O., Armstrong, R., Park, S.-M., Greenhalgh, L., Maye, U., Neill, N. J., Abbott, K. M., Sell, S., Ladda, R., Farber, D. M., Bader, P. I., Cushing, T., Drautz, J. M., Konczal, L., Nash, P., de Los Reyes, E., Carter, M. T., Hopkins, E., Marshall, C. R., Osborne, L. R., Gripp, K. W., Thrush, D. L., Hashimoto, S., Gastier-Foster, J. M., Astbury, C., Ylstra, B., Meijers-Heijboer, H., Posthuma, D., Menten, B., Mortier, G., Scherer, S. W., Eichler, E. E., Girirajan, S., Katsanis, N., Groffen, A. J., & Sistermans, E. A. (2013). Exonic deletions in AUTS2 cause a syndromic form of intellectual disability and suggest a critical role for the C terminus. *Am. J. Hum. Genet.*, 92(2), 210–220.

[14] Bianco, S. D. C., Vandepas, L., Correa-Medina, M., Gereben, B., Mukherjee, A., Kuohung, W., Carroll, R., Teles, M. G., Latronico, A. C., & Kaiser, U. B. (2011). KISS1R intracellular trafficking and degradation: effect of the Arg386Pro disease-associated mutation. *Endocrinology*, 152(4), 1616–1626.

[15] Bick, A. G., Flannick, J., Ito, K., Cheng, S., Vasan, R. S., Parfenov, M. G., Herman, D. S., DePalma, S. R., Gupta, N., Gabriel, S. B., Funke, B. H., Rehm, H. L., Benjamin, E. J., Aragam, J., Taylor Jr., H. A., Fox, E. R., Newton-Cheh, C., Kathiresan, S., O'Donnell, C. J., Wilson, J. G., Altshuler, D. M., Hirschhorn, J. N., Seidman, J. G., & Seidman, C. (2012). Burden of Rare Sarcomere Gene Variants in the Framingham and Jackson Heart Study Cohorts. *Am. J. Hum. Genet.*, 91(3), 513–519.

[16] Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., & Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421), 535–538.

[17] Bromberg, Y. & Rost, B. (2007). Snap: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res*, 35, 3823–35.

[18] Bromberg, Y., Yachdav, G., & Rost, B. (2008). SNAP predicts effect of mutations on protein function. *Bioinformatics*, 24(20), 2397–2398.

[19] Bromham, L. & Penny, D. (2003). The modern molecular clock. *Nat. Rev. Genet.*, 4(3), 216–224.

[20] Cai, W., Pei, J., & Grishin, N. V. (2004). Reconstruction of ancestral protein sequences and its applications. *BMC Evol Biol*, 4, 33.

[21] Canfield, M. C., Tamarappoo, B. K., Moses, A. M., Verkman, A. S., & Holtzman, E. J. (1997). Identification and characterization of aquaporin-2 water channel mutations causing nephrogenic diabetes insipidus with partial vasopressin response. *Hum. Mol. Genet.*, 6(11), 1865–1871.

[22] Capriotti, E., Calabrese, R., & Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22), 2729–2734.

[23] Carballo, S., Robinson, P., Otway, R., Fatkin, D., Jongbloed, J. D. H., de Jonge, N., Blair, E., van Tintelen, J. P., Redwood, C., & Watkins, H. (2009). Identification and functional characterization of cardiac troponin I as a novel disease gene in autosomal dominant dilated cardiomyopathy. *Circ. Res.*, 105(4), 375–382.

[24] Cariboni, A., Pimpinelli, F., Colamarino, S., Zaninetti, R., Piccolella, M., Rumio, C., Piva, F., Rugarli, E. I., & Maggi, R. (2004). The product of X-linked Kallmann's syndrome gene (KAL1) affects the migratory activity of gonadotropin-releasing hormone (GnRH)-producing neurons. *Hum. Mol. Genet.*, 13(22), 2781–2791.

[25] Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, 14 Suppl 3(Suppl 3), S3.

[26] Carvajal, C. A., Gonzalez, A. A., Romero, D. G., González, A., Mosso, L. M., Lagos, E. T., Hevia, M. d. P., Rosati, M. P., Perez-Acle, T. O., Gomez-Sanchez, C. E., Montero, J. A., & Fardella, C. E. (2003). Two homozygous mutations in the 11 beta-hydroxysteroid dehydrogenase type 2 gene in a case of apparent mineralocorticoid excess. *J. Clin. Endocrinol. Metab.*, 88(6), 2501–2507.

[27] Cassa, C. A., Tong, M. Y., & Jordan, D. M. (2013). Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum. Mutat.*, 34(9), 1216–1220.

[28] Chen, J. M., Raguenes, O., Ferec, C., Deprez, P. H., & Verellen-Dumoulin, C. (2000). A CGC>CAT gene conversion-like event resulting in the R122H mutation in the cationic

trypsinogen gene and its implication in the genotyping of pancreatitis. *J. Med. Genet.*, 37(11), E36.

[29] Cheng, J., Tester, D. J., Tan, B.-H., Valdivia, C. R., Kroboth, S., Ye, B., January, C. T., Ackerman, M. J., & Makielski, J. C. (2011). The common African American polymorphism SCN5A-S1103Y interacts with mutation SCN5A-R680H to increase late Na current. *Physiol. Genomics*, 43(9), 461–466.

[30] Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., & Fewell, G. A. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*.

[31] Chou, H.-H., Chiu, H.-C., Delaney, N. F., Segrè, D., & Marx, C. J. (2011). Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, 332(6034), 1190–1192.

[32] Chow, C. Y., Zhang, Y., Dowling, J. J., Jin, N., Adamska, M., Shiga, K., Szigeti, K., Shy, M. E., Li, J., Zhang, X., Lupski, J. R., Weisman, L. S., & Meisler, M. H. (2007). Mutation of FIG4 causes neurodegeneration in the pale tremor mouse and patients with CMT4J. *Nature*, 448(7149), 68–72.

[33] Chun, S. & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Res.*, 19(9), 1553–1561.

[34] Cooper, G. M. & Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.*, 12(9), 628–640.

[35] Corbett, M. A., Bahlo, M., Jolly, L., Afawi, Z., Gardner, A. E., Oliver, K. L., Tan, S., Coffey, A., Mulley, J. C., Dibbens, L. M., Simri, W., Shalata, A., Kivity, S., Jackson, G. D., Berkovic, S. F., & Gecz, J. (2010). A focal epilepsy and intellectual disability syndrome is due to a mutation in TBC1D24. *Am. J. Hum. Genet.*, 87(3), 371–375.

[36] Corbett-Detig, R. B., Zhou, J., Clark, A. G., Hartl, D. L., & Ayroles, J. F. (2013). Genetic incompatibilities are widespread within species. *Nature*, 504(7478), 135–137.

[37] Cosma, M. P., Pepe, S., Parenti, G., Settembre, C., Annunziata, I., Wade-Martins, R., Di Domenico, C., Di Natale, P., Mankad, A., Cox, B., Uziel, G., Mancini, G. M. S., Zammarchi, E., Donati, M. A., Kleijer, W. J., Filocamo, M., Carrozzo, R., Carella, M., & Ballabio, A. (2004). Molecular and functional analysis of SUMF1 mutations in multiple sulfatase deficiency. *Hum. Mutat.*, 23(6), 576–581.

[38] Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., Wheeler, D. A., Sabo, A., Lusk, C., Weiss, K. G., Akbar, H., Cree, A., Hawes, A. C., Newsham, I., Varghese, R. T., Villasana, D., Gross, S., Joshi, V., Santibanez, J., Morgan, M., Chang, K., Iv, W. H., Templeton, A. R., Boerwinkle,

E., Gibbs, R., & Sing, C. F. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun*, 1(8), 131.

[39] Crow, J. F. (2010). On epistasis: why it is unimportant in polygenic directional selection. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 365(1544), 1241–1244.

[40] Darbar, D., Kannankeril, P. J., Donahue, B. S., Kucera, G., Stubblefield, T., Haines, J. L., George, A. L., & Roden, D. M. (2008). Cardiac sodium channel (SCN5A) variants associated with atrial fibrillation. *Circulation*, 117(15), 1927–1935.

[41] Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, 6(12), e1001025.

[42] de Koning, A. P. J., Gu, W., Castoe, T. A., & Pollock, D. D. (2012). Phylogenetics, likelihood, evolution and complexity. *Bioinformatics*, 28(22), 2989–2990.

[43] De Oliveira Martins, L., Mallo, D., & Posada, D. (2014). A Bayesian Supertree Model for Genome-Wide Species Tree Reconstruction. *Systematic Biology*.

[44] Deen, P. M., Croes, H., van Aubel, R. A., Ginsel, L. A., & van Os, C. H. (1995). Water channels encoded by mutant aquaporin-2 genes in nephrogenic diabetes insipidus are impaired in their cellular routing. *J. Clin. Invest.*, 95(5), 2291–2296.

[45] Deen, P. M., Verdijk, M. A., Knoers, N. V., Wieringa, B., Monnens, L. A., van Os, C. H., & van Oost, B. A. (1994). Requirement of human renal water channel aquaporin-2 for vasopressin-dependent concentration of urine. *Science*, 264(5155), 92–95.

[46] Delport, W., Scheffler, K., Gravenor, M. B., Muse, S. V., & Kosakovsky Pond, S. (2010). Benchmarking multi-rate codon models. *PLoS ONE*, 5(7), e11587.

[47] DePristo, M. A., Weinreich, D. M., & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.*, 6(9), 678–687.

[48] Dinour, D., Gray, N. K., Ganon, L., Knox, A. J. S., Shalev, H., Sela, B.-A., Campbell, S., Sawyer, L., Shu, X., Valsamidou, E., Landau, D., Wright, A. F., & Holtzman, E. J. (2012). Two novel homozygous SLC2A9 mutations cause renal hypouricemia type 2. *Nephrol. Dial. Transplant.*, 27(3), 1035–1041.

[49] Dipple, K. M. & McCabe, E. R. B. (2000). Modifier Genes Convert "Simple" Mendelian Disorders to Complex Traits. *Molecular Genetics and Metabolism*, 71(1-2), 43–50.

[50] Dixon, M. E., Armstrong, P., Stevens, D. B., & Bamshad, M. (2001). Identical mutations in NOG can cause either tarsal/carpal coalition syndrome or proximal symphalangism. *Genet. Med.*, 3(5), 349–353.

[51] Dobbs, A. K., Yang, T., Farmer, D., Kager, L., Parolini, O., & Conley, M. E. (2007). Cutting edge: a hypomorphic mutation in Igbeta (CD79b) in a patient with immunodeficiency and a leaky defect in B cell development. *J. Immunol.*, 179(4), 2055–2059.

[52] Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2014). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*

[53] Dorfman, R., Nalpathamkalam, T., Taylor, C., Gonska, T., Keenan, K., Yuan, X. W., Corey, M., Tsui, L. C., Zielenski, J., & Durie, P. (2010). Do common in silico tools predict the clinical consequences of amino-acid substitutions in the cftr gene? *Clin Genet*, 77(5), 464–473.

[54] Fabrizi, G. M., Cavallaro, T., Angiari, C., Bertolasi, L., Cabrini, I., Ferrarini, M., & Rizzuto, N. (2004). Giant axon and neurofilament accumulation in Charcot-Marie-Tooth disease type 2E. *Neurology*, 62(8), 1429–1431.

[55] Ferrer-Costa, C., Orozco, M., & de la Cruz, X. (2007). Characterization of compensated mutations in terms of structural and physico-chemical properties. *J. Mol. Biol.*, 365(1), 249–256.

[56] Fitzgerald, T. W., Gerety, S. S., Jones, W. D., van Kogelenberg, M., King, D. A., McRae, J., Morley, K. I., Parthiban, V., Al-Turki, S., Ambridge, K., Barrett, D. M., Bayzetinova, T., Clayton, S., Coomber, E. L., Gribble, S., Jones, P., Krishnappa, N., Mason, L. E., Middleton, A., Miller, R., Prigmore, E., Rajan, D., Sifrim, A., Tivey, A. R., Ahmed, M., Akawi, N., Andrews, R., Anjum, U., Archer, H., Armstrong, R., Balasubramanian, M., Banerjee, R., Baralle, D., Batstone, P., Baty, D., Bennett, C., Berg, J., Bernhard, B., Bevan, A. P., Blair, E., Blyth, M., Bohanna, D., Bourdon, L., Bourn, D., Brady, A., Bragin, E., Brewer, C., Brueton, L., Brunstrom, K., Bumpstead, S. J., Bunyan, D. J., Burn, J., Burton, J., Canham, N., Castle, B., Chandler, K., Clasper, S., Clayton-Smith, J., Cole, T., Collins, A., Collinson, M. N., Connell, F., Cooper, N., Cox, H., Cresswell, L., Cross, G., Crow, Y., D'Alessandro, M., Dabir, T., Davidson, R., Davies, S., Dean, J., Deshpande, C., Devlin, G., Dixit, A., Dominiczak, A., Donnelly, C., Donnelly, D., Douglas, A., Duncan, A., Eason, J., Edkins, S., Ellard, S., Ellis, P., Elmslie, F., Evans, K., Everest, S., Fendick, T., Fisher, R., Flinter, F., Foulds, N., Fryer, A., Fu, B., Gardiner, C., Gaunt, L., Ghali, N., Gibbons, R., Gomes Pereira, S. L., Goodship, J., Goudie, D., Gray, E., Greene, P., Greenhalgh, L., Harrison, L., Hawkins, R., Hellens, S., Henderson, A., Hobson, E., Holden, S., Holder, S., Hollingsworth, G., Homfray, T., Humphreys, M., Hurst, J., Ingram, S., Irving, M., Jarvis, J., Jenkins, L., Johnson, D., Jones, D., Jones, E., Josifova, D., Joss, S., Kaemba, B., Kazembe, S., Kerr, B., Kini, U., Kinning, E., Kirby, G., Kirk, C., Kivuva, E., Kraus, A., Kumar, D., Lachlan, K., Lam, W., Lampe, A., Langman, C., Lees, M., Lim, D., Lowther, G., Lynch, S. A., Magee, A., Maher, E., Mansour, S., Marks, K., Martin, K., Maye, U., McCann, E., McConnell, V., McEntagart, M., McGowan, R., McKay, K., McKee, S., McMullan, D. J., McNerlan, S., Mehta, S., Metcalfe, K., Miles, E., Mohammed, S., Montgomery, T., Moore, D., Morgan, S., Morris, A., Morton, J.,

Mugalaasi, H., Murday, V., Nevitt, L., Newbury-Ecob, R., Norman, A., O'Shea, R., Ogilvie, C., Park, S., Parker, M. J., Patel, C., Paterson, J., Payne, S., Phipps, J., Pilz, D. T., Porteous, D., Pratt, N., Prescott, K., Price, S., Pridham, A., Procter, A., Purnell, H., Ragge, N., Rankin, J., Raymond, L., Rice, D., Robert, L., Roberts, E., Roberts, G., Roberts, J., Roberts, P., Ross, A., Rosser, E., Saggar, A., Samant, S., Sandford, R., Sarkar, A., Schweiger, S., Scott, C., Scott, R., Selby, A., Seller, A., Sequeira, C., Shannon, N., Sharif, S., Shaw-Smith, C., Shearing, E., Shears, D., Simonic, I., Simpkin, D., Singzon, R., Skitt, Z., Smith, A., Smith, B., Smith, K., Smithson, S., Sneddon, L., Splitt, M., Squires, M., Stewart, F., Stewart, H., Suri, M., Sutton, V., Swaminathan, G. J., Sweeney, E., Tatton-Brown, K., Taylor, C., Taylor, R., Tein, M., Temple, I. K., Thomson, J., Tolmie, J., Torokwa, A., Treacy, B., Turner, C., Turnpenny, P., Tysoe, C., Vandersteen, A., Vasudevan, P., Vogt, J., Wakeling, E., Walker, D., Waters, J., Weber, A., Wellesley, D., Whiteford, M., Widaa, S., Wilcox, S., Williams, D., Williams, N., Woods, G., Wragg, C., Wright, M., Yang, F., Yau, M., Carter, N. P., Parker, M., Firth, H. V., FitzPatrick, D. R., Wright, C. F., Barrett, J. C., & Hurles, M. E. (2014). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542), 223–228.

[57] Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadissa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T. J. P., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R., & Searle, S. M. J. (2013). Ensembl 2014. *Nucl. Acids Res.*, 42(D1), gkt1196–D755.

[58] Font, M. A., Feliubadaló, L., Estivill, X., Nunes, V., Golomb, E., Kreiss, Y., Pras, E., Bisceglia, L., d'Adamo, A. P., Zelante, L., Gasparini, P., Bassi, M. T., George, A. L., Manzoni, M., Riboni, M., Ballabio, A., Borsani, G., Reig, N., Fernández, E., Zorzano, A., Bertran, J., Palacín, M., & International Cystinuria Consortium (2001). Functional analysis of mutations in SLC7A9, and genotype-phenotype correlation in non-Type I cystinuria. *Hum. Mol. Genet.*, 10(4), 305–316.

[59] Fraïsse, C., Elderfield, J. A. D., & Welch, J. J. (2014). The genetics of speciation: are complex incompatibilities easier to evolve? *J. Evol. Biol.*, 27(4), 688–699.

[60] Gao, L. & Zhang, J. (2003). Why are some human disease-associated mutations fixed in mice? *Trends Genet.*, 19(12), 678–681.

[61] Geissler, W. M., Davis, D. L., Wu, L., Bradshaw, K. D., Patel, S., Mendonca, B. B., Elliston, K. O., Wilson, J. D., Russell, D. W., & Andersson, S. (1994). Male pseudohermaphroditism caused by mutations of testicular 17 beta-hydroxysteroid dehydrogenase 3. *Nat. Genet.*, 7(1), 34–39.

[62] Georgi, B., Voight, B. F., & Bućan, M. (2013). From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genet.*, 9(5), e1003484.

[63] Georgiou, D.-M., Zidar, J., Korosec, M., Middleton, L. T., Kyriakides, T., & Christodoulou, K. (2002). A novel NF-L mutation Pro22Ser is associated with CMT2 in a large Slovenian family. *Neurogenetics*, 4(2), 93–96.

[64] Georgopoulos, N. A., Pralong, F. P., Seidman, C. E., Seidman, J. G., Crowley, W. F., & Vallejo, M. (1997). Genetic heterogeneity evidenced by low incidence of KAL-1 gene mutations in sporadic cases of gonadotropin-releasing hormone deficiency. *J. Clin. Endocrinol. Metab.*, 82(1), 213–217.

[65] Germeshausen, M., Deerberg, S., Peter, Y., Reimer, C., Kratz, C. P., & Ballmaier, M. (2013). The spectrum of ELANE mutations and their implications in severe congenital and cyclic neutropenia. *Hum. Mutat.*, 34(6), 905–914.

[66] Gilbert-Dussardier, B., Segues, B., Rozet, J. M., Rabier, D., Calvas, P., de Lumley, L., Bonnefond, J. P., & Munnich, A. (1996). Partial duplication [dup. TCAC (178)] and novel point mutations (T125M, G188R, A209V, and H302L) of the ornithine transcarbamylase gene in congenital hyperammonemia. *Hum. Mutat.*, 8(1), 74–76.

[67] Gillespie, J. H. (2004). *Population Genetics*. Baltimore: Johns Hopkins University Press, second edition.

[68] Giudicessi, J. R. & Ackerman, M. J. (2013). Determinants of incomplete penetrance and variable expressivity in heritable cardiac arrhythmia syndromes. *Translational Research*, 161(1), 1–14.

[69] Goldgar, D. E., Easton, D. F., Deffenbaugh, A. M., Monteiro, A. N. A., Tavtigian, S. V., Couch, F. J., & BIC Steering Committee (2004). Integrated evaluation of dna sequence variants of unknown clinical significance: Application to brca1 and brca2. *Am J Hum Genet*, 75(4), 535–44.

[70] Gong, Y., Krakow, D., Marcelino, J., Wilkin, D., Chitayat, D., Babul-Hirji, R., Hudgins, L., Cremers, C. W., Cremers, F. P., Brunner, H. G., Reinker, K., Rimoin, D. L., Cohn, D. H., Goodman, F. R., Reardon, W., Patton, M., Francomano, C. A., & Warman, M. L. (1999). Heterozygous mutations in the gene encoding noggin affect human joint morphogenesis. *Nat. Genet.*, 21(3), 302–304.

[71] Gray, V. E., Kukurba, K. R., & Kumar, S. (2012). Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. *Bioinformatics*, 28(16), 2093–2096.

[72] Grimm, D. G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W., Duncan, L. E., & Borgwardt, K. M. (2015). The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Hum. Mutat.*, (pp. n/a–n/a).

[73] Grossi, S., Regis, S., Rosano, C., Corsolini, F., Uziel, G., Sessa, M., Di Rocco, M., Parenti, G., Deodato, F., Leuzzi, V., Biancheri, R., & Filocamo, M. (2008). Molecular analysis of ARSA and PSAP genes in twenty-one Italian patients with metachromatic leukodystrophy: identification and functional characterization of 11 novel ARSA alleles. *Hum. Mutat.*, 29(11), E220–30.

[74] Gulko, B., Hubisz, M. J., Gronau, I., & Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.*

[75] Guthrie, V. B., Allen, J., Camps, M., & Karchin, R. (2011). Network Models of TEM β-Lactamase Mutations Coevolving under Antibiotic Selection Show Modular Structure and Anticipate Evolutionary Trajectories. *PLoS Comput Biol*, 7(9), e1002184.

[76] Habart, D., Kalabova, D., Novotny, M., & Vorlova, Z. (2003). Thirty-four novel mutations detected in factor VIII gene by multiplex CSGE: modeling of 13 novel amino acid substitutions. *J. Thromb. Haemost.*, 1(4), 773–781.

[77] Haji-Seyed-Javadi, R., Jelodari-Mamaghani, S., Paylakhi, S. H., Yazdani, S., Nilforushan, N., Fan, J.-B., Klotzle, B., Mahmoudi, M. J., Ebrahimian, M. J., Chelich, N., Taghiabadi, E., Kamyab, K., Boileau, C., Paisan-Ruiz, C., Ronaghi, M., & Elahi, E. (2012). LTBP2 mutations cause Weill-Marchesani and Weill-Marchesani-like syndrome and affect disruptions in the extracellular matrix. *Hum. Mutat.*, 33(8), 1182–1187.

[78] Halonen, M., Kangas, H., Rüppell, T., Ilmarinen, T., Ollila, J., Kolmer, M., Vihinen, M., Palvimo, J., Saarela, J., Ulmanen, I., & Eskelin, P. (2004). APECED-causing mutations in AIRE reveal the functional domains of the protein. *Hum. Mutat.*, 23(3), 245–257.

[79] Hardison, R. C. & Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.*, 13(7), 469–483.

[80] Henikoff, J. G. & Henikoff, S. (1996). Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci*, 12(2), 135–143.

[81] Hill, W. G., Goddard, M. E., & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.*, 4(2), e1000008.

[82] Himmel, D. M., Gourinath, S., Reshetnikova, L., Shen, Y., Szent-Györgi, A. G., & Cohen, C. (2002). Crystallographic findings on the internally uncoupled and near-rigor states of myosin: further insights into the mechanics of the motor. *Proc Natl Acad Sci USA*, 99, 12645–50.

[83] Homer & Fagles, R. (1990). *The Iliad*. New York, N.Y., U.S.A.: Penguin Books.

[84] Homer & Fagles, R. (1996). *The Odyssey*. New York, N.Y., U.S.A.: Penguin Books.

[85] Horvath, A., Giatzakis, C., Tsang, K., Greene, E., Osorio, P., Boikos, S., Libè, R., Patronas, Y., Robinson-White, A., Remmers, E., Bertherat, J., Nesterova, M., & Stratakis, C. A. (2008a). A cAMP-specific phosphodiesterase (PDE8B) that is mutated in adrenal hyperplasia is expressed widely in human and mouse tissues: a novel PDE8B isoform in human adrenal cortex. *Eur. J. Hum. Genet.*, 16(10), 1245–1253.

[86] Horvath, A., Mericq, V., & Stratakis, C. A. (2008b). Mutation in PDE8B, a cyclic AMP-specific phosphodiesterase in adrenal hyperplasia. *N. Engl. J. Med.*, 358(7), 750–752.

[87] Houdusse, A., Kalabokis, V. N., Himmel, D., Szent-Györgi, A. G., & Cohen, C. (1999). Atomic structure of scallop myosin subfragment s1 complexed with mgadp: a novel conformation of the myosin head. *Cell*, 97, 459–70.

[88] Hsi, G., Cullen, L. M., Macintyre, G., Chen, M. M., Glerum, D. M., & Cox, D. W. (2008). Sequence variation in the ATP-binding domain of the Wilson disease transporter, ATP7B, affects copper transport in a yeast model system. *Hum. Mutat.*, 29(4), 491–501.

[89] Hu, Y., Guimond, S. E., Travers, P., Cadman, S., Hohenester, E., Turnbull, J. E., Kim, S.-H., & Bouloux, P.-M. (2009). Novel mechanisms of fibroblast growth factor receptor 1 regulation by extracellular matrix protein anosmin-1. *Journal of Biological Chemistry*, 284(43), 29905–29920.

[90] Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protoc.*, 4(1), 44–57.

[91] Huang, W., Richards, S., Carbone, M. A., Zhu, D., Anholt, R. R. H., Ayroles, J. F., Duncan, L., Jordan, K. W., Lawrence, F., Magwire, M. M., Warner, C. B., Blankenburg, K., Han, Y., Javaid, M., Jayaseelan, J., Jhangiani, S. N., Muzny, D., Ongeri, F., Perales, L., Wu, Y.-Q., Zhang, Y., Zou, X., Stone, E. A., Gibbs, R. A., & Mackay, T. F. C. (2012). Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proc. Natl. Acad. Sci. U.S.A.*, 109(39), 15553–15559.

[92] Huelsenbeck, J. P., Joyce, P., Lakner, C., & Ronquist, F. (2008). Bayesian analysis of amino acid substitution models. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 363(1512), 3941–3953.

[93] Huzurbazar, S., Kolesov, G., Massey, S. E., Harris, K. C., Churbanov, A., & Liberles, D. A. (2010). Lineage-Specific Differences in the Amino Acid Substitution Process. *J. Mol. Biol.*, 396(5), 1410–1421.

[94] Hwu, W.-L., Chien, Y.-H., Lee, N.-C., Chiang, S.-C., Dobrovolny, R., Huang, A.-C., Yeh, H.-Y., Chao, M.-C., Lin, S.-J., Kitagawa, T., Desnick, R. J., & Hsu, L.-W. (2009). Newborn

screening for Fabry disease in Taiwan reveals a high incidence of the later-onset GLA mutation c.936+919G>A (IVS4+919G>A). *Hum. Mutat.*, 30(10), 1397–1405.

[95] Isbrandt, D., Arlt, G., Brooks, D. A., Hopwood, J. J., von Figura, K., & Peters, C. (1994). Mucopolysaccharidosis VI (Maroteaux-Lamy syndrome): six unique arylsulfatase B gene alleles causing variable disease phenotypes. *Am. J. Hum. Genet.*, 54(3), 454–463.

[96] Iseri, S. U., Osborne, R. J., Farrall, M., Wyatt, A. W., Mirza, G., Nürnberg, G., Kluck, C., Herbert, H., Martin, A., Hussain, M. S., Collin, J. R. O., Lathrop, M., Nürnberg, P., Ragoussis, J., & Ragge, N. K. (2009). Seeing clearly: the dominant and recessive nature of FOXE3 in eye developmental anomalies. *Hum. Mutat.*, 30(10), 1378–1386.

[97] Jayo, A., Pabón, D., Lastres, P., Jiménez-Yuste, V., & González-Manchón, C. (2006). Type II Glanzmann thrombasthenia in a compound heterozygote for the alpha IIb gene. A novel missense mutation in exon 27. *Haematologica*, 91(10), 1352–1359.

[98] Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8(3), 275–282.

[99] Jordan, D. M., Frangakis, S. G., Golzio, C., Cassa, C., Kurtzberg, J., for Neonatal Genomics, T. F., Davis, E. E., Sunyaev, S. R., & Katsanis, N. (in press). Identification of cis-suppression of human disease mutations by comparative genomics. *Nature*.

[100] Jordan, D. M., Kiezun, A., Baxter, S. M., Agarwala, V., Green, R. C., Murray, M. F., Pugh, T., Lebo, M. S., Rehm, H. L., Funke, B. H., & Sunyaev, S. R. (2011). Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. *Am. J. Hum. Genet.*, 88(2), 183–192.

[101] Jordan, D. M., Ramensky, V. E., & Sunyaev, S. R. (2010). Human allelic variation: perspective from protein function, structure, and evolution. *Curr. Opin. Struct. Biol.*, 20(3), 342–350.

[102] Kaname, T., Yanagi, K., Chinen, Y., Makita, Y., Okamoto, N., Maehara, H., Owan, I., Kanaya, F., Kubota, Y., Oike, Y., Yamamoto, T., Kurosawa, K., Fukushima, Y., Bohring, A., Opitz, J. M., Yoshiura, K.-I., Niikawa, N., & Naritomi, K. (2007). Mutations in CD96, a member of the immunoglobulin superfamily, cause a form of the C (Opitz trigonocephaly) syndrome. *Am. J. Hum. Genet.*, 81(4), 835–841.

[103] Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hinrichs, A. S., Learned, K., Lee, B. T., Li, C. H., Raney, B. J., Rhead, B., Rosenbloom, K. R., Sloan, C. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., & Kent, W. J. (2014). The UCSC Genome Browser database: 2014 update. *Nucl. Acids Res.*, 42(1), D764–70.

[104] Katsanis, N., Cotten, M., & Angrist, M. (2012). *Exome and genome sequencing of neonates with neurodevelopmental disorders.* Future Neurology.

[105] Katsanis, N., Eichers, E. R., Ansley, S. J., Lewis, R. A., Kayserili, H., Hoskins, B. E., Scambler, P. J., Beales, P. L., & Lupski, J. R. (2002). BBS4 is a minor contributor to Bardet-Biedl syndrome and may also participate in triallelic inheritance. *Am. J. Hum. Genet.*, 71(1), 22–29.

[106] Khanna, H., Davis, E. E., Murga-Zamalloa, C. A., Estrada-Cuzcano, A., Lopez, I., den Hollander, A. I., Zonneveld, M. N., Othman, M. I., Waseem, N., Chakarova, C. F., Maubaret, C., Diaz-Font, A., MacDonald, I., Muzny, D. M., Wheeler, D. A., Morgan, M., Lewis, L. R., Logan, C. V., Tan, P. L., Beer, M. A., Inglehearn, C. F., Lewis, R. A., Jacobson, S. G., Bergmann, C., Beales, P. L., Attié-Bitach, T., Johnson, C. A., Otto, E. A., Bhattacharya, S. S., Hildebrandt, F., Gibbs, R. A., Koenekoop, R. K., Swaroop, A., & Katsanis, N. (2009). A common allele in RPGRIP1L is a modifier of retinal degeneration in ciliopathies. *Nat. Genet.*, 41(6), 739–745.

[107] Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C. J., Binder, J. X., Malone, J., Vasant, D., Parkinson, H., & Schriml, L. M. (2015). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucl. Acids Res.*, 43(Database issue), D1071–8.

[108] Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, 46(3), 310–315.

[109] Kleefstra, T., Yntema, H. G., Oudakker, A. R., Banning, M. J. G., Kalscheuer, V. M., Chelly, J., Moraine, C., Ropers, H.-H., Fryns, J.-P., Janssen, I. M., Sistermans, E. A., Nillesen, W. N., de Vries, L. B. A., Hamel, B. C. J., & van Bokhoven, H. (2004). Zinc finger 81 (ZNF81) mutations associated with X-linked mental retardation. *J. Med. Genet.*, 41(5), 394–399.

[110] Kogawa, K., Kudoh, J., Nagafuchi, S., Ohga, S., Katsuta, H., Ishibashi, H., Harada, M., Hara, T., & Shimizu, N. (2002). Distinct clinical phenotype and immunoreactivity in Japanese siblings with autoimmune polyglandular syndrome type 1 (APS-1) associated with compound heterozygous novel AIRE gene mutations. *Clin. Immunol.*, 103(3 Pt 1), 277–283.

[111] Kohane, I. S., Hsing, M., & Kong, S. W. (2012). Taxonomizing, sizing, and overcoming the incidentalome. *Genet. Med.*, 14(4), 399–404.

[112] Kondrashov, A. S., Sunyaev, S., & Kondrashov, F. A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 99(23), 14878–14883.

[113] Kousi, M., Lehesjoki, A.-E., & Mole, S. E. (2012). Update of the mutation spectrum and clinical correlations of over 360 mutations in eight genes that underlie the neuronal ceroid lipofuscinoses. *Hum. Mutat.*, 33(1), 42–63.

[114] Kremeyer, B., Lopera, F., Cox, J. J., Momin, A., Rugiero, F., Marsh, S., Woods, C. G., Jones, N. G., Paterson, K. J., Fricker, F. R., Villegas, A., Acosta, N., Pineda-Trujillo, N. G., Ramírez, J. D., Zea, J., Burley, M.-W., Bedoya, G., Bennett, D. L. H., Wood, J. N., & Ruiz-Linares, A. (2010). A gain-of-function mutation in TRPA1 causes familial episodic pain syndrome. *Neuron*, 66(5), 671–680.

[115] Kulathinal, R. J., Bettencourt, B. R., & Hartl, D. L. (2004). Compensated deleterious mutations in insect genomes. *Science*, 306(5701), 1553–1554.

[116] Lacombe, A., Lee, H., Zahed, L., Choucair, M., Muller, J.-M., Nelson, S. F., Salameh, W., & Vilain, E. (2006). Disruption of POF1B binding to nonmuscle actin filaments is associated with premature ovarian failure. *Am. J. Hum. Genet.*, 79(1), 113–119.

[117] Landouré, G., Zdebik, A. A., Martinez, T. L., Burnett, B. G., Stanescu, H. C., Inada, H., Shi, Y., Taye, A. A., Kong, L., Munns, C. H., Choo, S. S., Phelps, C. B., Paudel, R., Houlden, H., Ludlow, C. L., Caterina, M. J., Gaudet, R., Kleta, R., Fischbeck, K. H., & Sumner, C. J. (2010). Mutations in TRPV4 cause Charcot-Marie-Tooth disease type 2C. *Nat. Genet.*, 42(2), 170–174.

[118] Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucl. Acids Res.*, 42(Database issue), D980–5.

[119] Lenk, G. M., Ferguson, C. J., Chow, C. Y., Jin, N., Jones, J. M., Grant, A. E., Zolov, S. N., Winters, J. J., Giger, R. J., Dowling, J. J., Weisman, L. S., & Meisler, M. H. (2011). Pathogenic mechanism of the FIG4 mutation responsible for Charcot-Marie-Tooth disease CMT4J. *PLoS Genet.*, 7(6), e1002104.

[120] Letunic, I., Doerks, T., & Bork, P. (2008). Smart 6: recent updates and new developments. *Nucleic Acids Res*, 36(Database issue), D229–32.

[121] Li, C., Liakata, M., & Rebholz-Schuhmann, D. (2014). Biological network extraction from scientific literature: state of the art and challenges. *Briefings in Bioinformatics*, 15(5), 856–877.

[122] Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Broad Institute Sequencing Platform, Team, W. G. A., Worley, K. C., Kovar, C. L., Muzny, D. M., Gibbs, R. A., Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Warren, W. C., Mardis, E. R., Weinstock, G. M., Wilson, R. K., Genome Institute at Washington University, Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D., Siepel, A., Goldman, N., Pollard,

K. S., Pedersen, J. S., Lander, E. S., & Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370), 476–482.

[123] Liu, X., Jian, X., & Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, 34(9), E2393–402.

[124] Lunzer, M., Golding, G. B., & Dean, A. M. (2010). Pervasive cryptic epistasis in molecular evolution. *PLoS Genet.*, 6(10), e1001162.

[125] Lupas, A. (1996). Prediction and analysis of coiled-coil structures. *Methods Enzymol*, 266, 513–525.

[126] Lupas, A., van Dyke, M., & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science*, 252, 1162–4.

[127] Maret, A., Ding, C., Kornfield, S. L., & Levine, M. A. (2008). Analysis of the GCM2 gene in isolated hypoparathyroidism: a molecular and biochemical study. *J. Clin. Endocrinol. Metab.*, 93(4), 1426–1432.

[128] Marston, S., Copeland, O., Jacques, A., Livesey, K., Tsang, V., McKenna, W. J., Jalilzadeh, S., Carballo, S., Redwood, C., & Watkins, H. (2009). Evidence from human myectomy samples that mybpc3 mutations cause hypertrophic cardiomyopathy through haploinsufficiency. *Circ Res*, 105(3), 219–222.

[129] Martinez, L., Andreani, R., & Martinez, J. M. (2007). Convergent algorithms for protein structural alignment. *BMC Bioinformatics*, 8, 306.

[130] McCandlish, D. M., Rajon, E., Shah, P., Ding, Y., & Plotkin, J. B. (2013). The role of epistasis in protein evolution. *Nature*, 497(7451), E1–E2.

[131] Menassa, R., Tardy, V., Despert, F., Bouvattier-Morel, C., Brossier, J. P., Cartigny, M., & Morel, Y. (2008). p.H62L, a rare mutation of the CYP21 gene identified in two forms of 21-hydroxylase deficiency. *J. Clin. Endocrinol. Metab.*, 93(5), 1901–1908.

[132] Messika-Zeitoun, L., Gouédard, L., Belville, C., Dutertre, M., Lins, L., Imbeaud, S., Hughes, I. A., Picard, J. Y., Josso, N., & di Clemente, N. (2001). Autosomal recessive segregation of a truncating mutation of anti-Müllerian type II receptor in a family affected by the persistent Müllerian duct syndrome contrasts with its dominant negative activity in vitro. *J. Clin. Endocrinol. Metab.*, 86(9), 4390–4397.

[133] Mestroni, L. & Taylor, M. R. G. (2013). Genetics and genetic testing of dilated cardiomyopathy: a new perspective. *Discov Med*, 15(80), 43–49.

[134] Millar, D. S., Lewis, M. D., Horan, M., Newsway, V., Easter, T. E., Gregory, J. W., Fryklund, L., Norin, M., Crowne, E. C., Davies, S. J., Edwards, P., Kirk, J., Waldron, K., Smith, P. J., Phillips, J. A., Scanlon, M. F., Krawczak, M., Cooper, D. N., & Procter, A. M. (2003). Novel mutations of the growth hormone 1 (GH1) gene disclosed by modulation of the clinical selection criteria for individuals with short stature. *Hum. Mutat.*, 21(4), 424–440.

[135] Mohapatra, B., Casey, B., Li, H., Ho-Dawson, T., Smith, L., Fernbach, S. D., Molinari, L., Niesh, S. R., Jefferies, J. L., Craigen, W. J., Towbin, J. A., Belmont, J. W., & Ware, S. M. (2009). Identification and functional characterization of NODAL rare variants in heterotaxy and isolated cardiovascular malformations. *Hum. Mol. Genet.*, 18(5), 861–871.

[136] Molatore, S., Russo, M. T., D'Agostino, V. G., Barone, F., Matsumoto, Y., Albertini, A. M., Minoprio, A., Degan, P., Mazzei, F., Bignami, M., & Ranzani, G. N. (2010). MUTYH mutations associated with familial adenomatous polyposis: functional characterization by a mammalian cell-based assay. *Hum. Mutat.*, 31(2), 159–166.

[137] Molinski, S. V., Gonska, T., Huan, L. J., Baskin, B., Janahi, I. A., Ray, P. N., & Bear, C. E. (2014). Genetic, cell biological, and clinical interrogation of the CFTR mutation c.3700 A>G (p.Ile1234Val) informs strategies for future medical intervention. *Genet. Med.*, 16(8), 625–632.

[138] Montagnoli, A., Guardavaccaro, D., Starace, G., & Tirone, F. (1996). Overexpression of the nerve growth factor-inducible PC3 immediate early gene is associated with growth inhibition. *Cell Growth Differ.*, 7(10), 1327–1336.

[139] Mottaz, A., David, F. P. A., Veuthey, A.-L., & Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, 26(6), 851–852.

[140] Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.*, 30(5), 1196–1205.

[141] Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., Woollard, P., Topp, S., Hall, M. D., Nangle, K., Wang, J., Abecasis, G., Cardon, L. R., Zöllner, S., Whittaker, J. C., Chissoe, S. L., Novembre, J., & Mooser, V. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090), 100–104.

[142] Ng, P. C. & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.*, 11(5), 863–874.

[143] Ng, P. C. & Henikoff, S. (2003). Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31, 3812–4.

[144] Ng, P. C. & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*, 7, 61–80.

[145] Nielsen, R. (2005). Molecular signatures of natural selection. *Annu Rev Genet*.

[146] Nikoshkov, A., Lajic, S., Holst, M., Wedell, A., & Luthman, H. (1997). Synergistic effect of partially inactivating mutations in steroid 21-hydroxylase deficiency. *J. Clin. Endocrinol. Metab.*, 82(1), 194–199.

[147] Ohlsson, G., Müller, J., Skakkebaek, N. E., & Schwartz, M. (1999). Steroid 21-hydroxylase deficiency: mutational spectrum in Denmark, three novel mutations, and in vitro expression analysis. *Hum. Mutat.*, 13(6), 482–486.

[148] Ohno, K., Engel, A. G., Brengman, J. M., Shen, X. M., Heidenreich, F., Vincent, A., Milone, M., Tan, E., Demirci, M., Walsh, P., Nakano, S., & Akiguchi, I. (2000). The spectrum of mutations causing end-plate acetylcholinesterase deficiency. *Ann. Neurol.*, 47(2), 162–170.

[149] Padovano, V., Lucibello, I., Alari, V., Della Mina, P., Crespi, A., Ferrari, I., Recagni, M., Lattuada, D., Righi, M., Toniolo, D., Villa, A., & Pietrini, G. (2011). The POF1B candidate gene for premature ovarian failure regulates epithelial polarity. *J. Cell. Sci.*, 124(Pt 19), 3356–3368.

[150] Pahl, S., Pavlova, A., Driesen, J., & Oldenburg, J. (2014). Effect of F8 B domain gene variants on synthesis, secretion, activity and stability of factor VIII protein. *Thromb. Haemost.*, 111(1), 58–66.

[151] Pfaeffle, R. W., Hunter, C. S., Savage, J. J., Duran-Prado, M., Mullen, R. D., Neeb, Z. P., Eiholzer, U., Hesse, V., Haddad, N. G., Stobbe, H. M., Blum, W. F., Weigel, J. F. W., & Rhodes, S. J. (2008). Three novel missense mutations within the LHX4 gene are associated with variable pituitary hormone deficiencies. *J. Clin. Endocrinol. Metab.*, 93(3), 1062–1071.

[152] Pfützer, R., Myers, E., Applebaum-Shapiro, S., Finch, R., Ellis, I., Neoptolemos, J., Kant, J. A., & Whitcomb, D. C. (2002). Novel cationic trypsinogen (PRSS1) N29T and R122C mutations cause autosomal dominant hereditary pancreatitis. *Gut*, 50(2), 271–272.

[153] Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., Greenblatt, M. S., Hogervorst, F. B., Hoogerbrugge, N., Spurdle, A. B., Tavtigian, S. V., & IARC Unclassified Genetic Variants Working Group (2008). Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat*, 29(11), 1282–1291.

[154] Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20(1), 110–121.

[155] Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5), 676–679.

[156] Poon, A., Davis, B. H., & Chao, L. (2005). The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. *Genetics*, 170(3), 1323–1332.

[157] Rezaei, N., Moin, M., Pourpak, Z., Ramyar, A., Izadyar, M., Chavoshzadeh, Z., Sherkat, R., Aghamohammadi, A., Yeganeh, M., Mahmoudi, M., Mahjoub, F., Germeshausen, M., Grudzien, M., Horwitz, M. S., Klein, C., & Farhoudi, A. (2007). The clinical, immunohematological, and molecular study of Iranian patients with severe congenital neutropenia. *J. Clin. Immunol.*, 27(5), 525–533.

[158] Richard, P., Charron, P., Carrier, L., Ledeuil, C., Cheav, T., Pichereau, C., Benaiche, A., Isnard, R., Dubourg, O., Burban, M., Gueffet, J. P., Millaire, A., Desnos, M., Schwartz, K., Hainque, B., & Komajda, M. (2003). Hypertrophic cardiomyopathy: distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy. *Circulation*, 107, 2227–2232.

[159] Richards, C. S., Bale, S., Bellissimo, D. B., Das, S., Grody, W. W., Hegde, M. R., Lyon, E., Ward, B. E., & Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee (2008). Acmg recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet Med*, 10(4), 294–300.

[160] Rohmann, E., Brunner, H. G., Kayserili, H., Uyguner, O., Nürnberg, G., Lew, E. D., Dobbie, A., Eswarakumar, V. P., Uzumcu, A., Ulubil-Emeroglu, M., Leroy, J. G., Li, Y., Becker, C., Lehnerdt, K., Cremers, C. W. R. J., Yüksel-Apak, M., Nürnberg, P., Kubisch, C., Kubisch, C., Schlessinger, J., van Bokhoven, H., & Wollnik, B. (2006). Mutations in different components of FGF signaling in LADD syndrome. *Nat. Genet.*, 38(4), 414–417.

[161] Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*.

[162] Sakuraba, H., Oshima, A., Fukuhara, Y., Shimmoto, M., Nagao, Y., Bishop, D. F., Desnick, R. J., & Suzuki, Y. (1990). Identification of point mutations in the alpha-galactosidase A gene in classical and atypical hemizygotes with Fabry disease. *Am. J. Hum. Genet.*, 47(5), 784–789.

[163] Salipante, S. J., Benson, K. F., Luty, J., Hadavi, V., Kariminejad, R., Kariminejad, M. H., Rezaei, N., & Horwitz, M. S. (2007). Double de novo mutations of ELA2 in cyclic and severe congenital neutropenia. *Hum. Mutat.*, 28(9), 874–881.

[164] Sasaki, T., Gotow, T., Shiozaki, M., Sakaue, F., Saito, T., Julien, J.-P., Uchiyama, Y., & Hisanaga, S.-I. (2006). Aggregate formation and phosphorylation of neurofilament-L Pro22 Charcot-Marie-Tooth disease mutants. *Hum. Mol. Genet.*, 15(6), 943–952.

[165] Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucl. Acids Res.*, 40(Web Server issue), W452–7.

[166] Simon, P., Weiss, F. U., Sahin-Toth, M., Parry, M., Nayler, O., Lenfers, B., Schnekenburger, J., Mayerle, J., Domschke, W., & Lerch, M. M. (2002). Hereditary pancreatitis caused by a novel PRSS1 mutation (Arg-122 –> Cys) that alters autoactivation and autodegradation of cationic trypsinogen. *J. Biol. Chem.*, 277(7), 5404–5410.

[167] Soylemez, O. & Kondrashov, F. A. (2012). Estimating the rate of irreversibility in protein evolution. *Genome Biol Evol*, 4(12), 1213–1222.

[168] Speiser, P. W., Dupont, J., Zhu, D., Serrat, J., Buegeleisen, M., Tusie-Luna, M. T., Lesser, M., New, M. I., & White, P. C. (1992). Disease expression and molecular genotype in congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *J. Clin. Invest.*, 90(2), 584–595.

[169] Spodsberg, N., Jacob, R., Alfalah, M., Zimmer, K. P., & Naim, H. Y. (2001). Molecular basis of aberrant apical protein transport in an intestinal enzyme disorder. *J. Biol. Chem.*, 276(26), 23506–23510.

[170] Studer, R. A., Dessailly, B. H., & Orengo, C. A. (2013). Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem. J.*, 449(3), 581–594.

[171] Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., & Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, 12(5), 387–394.

[172] Sunyaev, S. R., Ramensky, V. E., Koch, I., Lathe, III, W., Kondrashov, A. S., & Bork, P. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.*, 10(6), 591–597.

[173] Suriano, G., Azevedo, L., Novais, M., Boscolo, B., Seruca, R., Amorim, A., & Ghibaudi, E. M. (2007). In vitro demonstration of intra-locus compensation using the ornithine transcarbamylase protein as model. *Hum. Mol. Genet.*, 16(18), 2209–2214.

[174] Tamura, K. & Kumar, S. (2002). Evolutionary Distance Estimation Under Heterogeneous Substitution Pattern Among Lineages. *Mol. Biol. Evol.*

[175] Tavtigian, S. V., Greenblatt, M. S., Lesueur, F., Byrnes, G. B., & IARC Unclassified Genetic Variants Working Group (2008). In silico analysis of missense substitutions using sequence-alignment based methods. *Hum Mutat*, 29(11), 1327–1336.

[176] Tchernitchko, D., Goossens, M., & Wajcman, H. (2004). In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. *Clin Chem*, 50(11), 1974–1978.

[177] Teich, N., Bauer, N., Mössner, J., & Keim, V. (2002). Mutational screening of patients with nonalcoholic chronic pancreatitis: identification of further trypsinogen variants. *Am. J. Gastroenterol.*, 97(2), 341–346.

[178] Teles, M. G., Bianco, S. D. C., Brito, V. N., Trarbach, E. B., Kuohung, W., Xu, S., Seminara, S. B., Mendonca, B. B., Kaiser, U. B., & Latronico, A. C. (2008). A GPR54-activating mutation in a patient with central precocious puberty. *N. Engl. J. Med.*, 358(7), 709–715.

[179] Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., Broad GO, Seattle GO, & NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090), 64–69.

[180] Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., & Narechania, A. (2003). Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13, 2129–2141.

[181] Thomas, P. D., Kejariwal, A., Guo, N., Mi, H., Campbell, M. J., Muruganujan, A., & Lazareva-Ulitsky, B. (1 July 2006). Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res*, 34(suppl 2), W645–W650.

[182] Thomée, C., Schubert, S. W., Parma, J., Lê, P. Q., Hashemolhosseini, S., Wegner, M., & Abramowicz, M. J. (2005). GCMB mutation in familial isolated hypoparathyroidism with residual secretion of parathyroid hormone. *J. Clin. Endocrinol. Metab.*, 90(5), 2487–2492.

[183] Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, 32(4), 358–368.

[184] Thusberg, J. & Vihinen, M. (2009). Pathogenic or not? and if so, then how? studying the effects of missense mutations using bioinformatics methods. *Hum Mutat*, 30(5), 703–714.

[185] Tusie-Luna, M. T., Speiser, P. W., Dumic, M., New, M. I., & White, P. C. (1991). A mutation (Pro-30 to Leu) in CYP21 represents a potential nonclassic steroid 21-hydroxylase deficiency allele. *Mol. Endocrinol.*, 5(5), 685–692.

[186] Vargas-Poussou, R., Forestier, L., Dautzenberg, M. D., Niaudet, P., Déchaux, M., & Antignac, C. (1997). Mutations in the vasopressin V2 receptor and aquaporin-2 genes in 12 families with congenital nephrogenic diabetes insipidus. *J. Am. Soc. Nephrol.*, 8(12), 1855–1862.

[187] Vieira, T. C., Dias da Silva, M. R., Cerutti, J. M., Brunner, E., Borges, M., Arnaldi, L. T., Kopp, P., & Abucham, J. (2003). Familial combined pituitary hormone deficiency due to a novel mutation R99Q in the hot spot region of Prophet of Pit-1 presenting as constitutional growth delay. *J. Clin. Endocrinol. Metab.*, 88(1), 38–44.

[188] Vinogradova, M. V., Stone, D. B., Malanina, G. G., Karatzaferi, C., Cooke, R., Medelson, R. A., & Fletterick, R. J. (2005). Ca(2+)-regulated structural changes in troponin. *Proc Natl Acad Sci USA*, 102, 5038–43.

[189] Wang, D. W., Desai, R. R., Crotti, L., Arnestad, M., Insolia, R., Pedrazzini, M., Ferrandi, C., Vege, A., Rognum, T., Schwartz, P. J., & George, A. L. (2007). Cardiac sodium channel dysfunction in sudden infant death syndrome. *Circulation*, 115(3), 368–376.

[190] Wang, L., Seidman, J. G., & Seidman, C. E. (2010). Narrative review: harnessing molecular genetics for the diagnosis and management of hypertrophic cardiomyopathy. *Ann Intern Med*, 152, 513–20.

[191] Wang, L., Wang, L., He, F., Bu, J., Zhen, Y., Liu, X., Liu, X., Du, W., Dong, J., Cooney, J. D., Dubey, S. K., Shi, Y., Gong, B., Li, J., McBride, P. F., Jia, Y., Lu, F., Soltis, K. A., Lin, Y., Namburi, P., Liang, C., Sundaresan, P., Paw, B. H., Li, W., Li, D. Y., Phillips, J. D., & Yang, Z. (2012). ABCB6 mutations cause ocular coloboma. *Am. J. Hum. Genet.*, 90(1), 40–48.

[192] Weinreich, D. M., Delaney, N. F., DePristo, M. A., & Hartl, D. L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770), 111–114.

[193] Weinreich, D. M., Lan, Y., Wylie, C. S., & Heckendorn, R. B. (2013). Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.*, 23(6), 700–707.

[194] Whitcomb, D. C., Gorry, M. C., Preston, R. A., Furey, W., Sossenheimer, M. J., Ulrich, C. D., Martin, S. P., Gates, L. K., Amann, S. T., Toskes, P. P., Liddle, R., McGrath, K., Uomo, G., Post, J. C., & Ehrlich, G. D. (1996). Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. *Nat. Genet.*, 14(2), 141–145.

[195] Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M., & Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Res.*

[196] Wieland, I., Sabathil, J., Ostendorf, A., Rittinger, O., Röpke, A., Winnepenninckx, B., Kooy, F., Holinski-Feder, E., & Wieacker, P. (2005). A missense mutation in the coiled-coil motif of the HP1-interacting domain of ATR-X in a family with X-linked mental retardation. *Neurogenetics*, 6(1), 45–47.

[197] Witt, H., Luck, W., & Becker, M. (1999). A signal peptide cleavage site mutation in the cationic trypsinogen gene is strongly associated with chronic pancreatitis. *Gastroenterology*, 117(1), 7–10.

[198] Wright, A., Charlesworth, B., Rudan, I., Carothers, A., & Campbell, H. (2003). A polygenic basis for late-onset disease. *Trends in Genetics*, 19(2), 97–106.

[199] Wu, J. Y., Yang, C. F., Lee, C. C., Chang, J. G., & Tsai, F. J. (2000). A novel mutation (Q239R) identified in a Taiwan Chinese patient with type VI mucopolysaccharidosis (Maroteaux-Lamy syndrome). *Hum. Mutat.*, 15(4), 389–390.

[200] Yang, Y., Li, J., Lin, X., Yang, Y., Hong, K., Wang, L., Liu, J., Li, L., Yan, D., Liang, D., Xiao, J., Jin, H., Wu, J., Zhang, Y., & Chen, Y.-H. (2009). Novel KCNA5 loss-of-function mutations responsible for atrial fibrillation. *J. Hum. Genet.*, 54(5), 277–283.

[201] Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*.

[202] Yu, Y., Dong, J., Liu, K. J., & Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci. U.S.A.*, 111(46), 16448–16453.

[203] Yue, P., Melamud, E., & Moult, J. (2006). Snps3d: candidate gene and snp selection for association studies. *BMC Bioinformatics*, 7, 166.

[204] Yue, P. & Moult, J. (2006). Identification and analysis of deleterious human snps. *J Mol Biol*, 356, 1263–74.

[205] Zaghloul, N. A. & Katsanis, N. (2010). Functional modules, mutational load and human genetic disease. *Trends in Genetics*, 26(4), 168–176.

[206] Zaghloul, N. A., Liu, Y., Gerdes, J. M., Gascue, C., Oh, E. C., Leitch, C. C., Bromberg, Y., Binkley, J., Leibel, R. L., Sidow, A., Badano, J. L., & Katsanis, N. (2010). Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. *Proc. Natl. Acad. Sci. U.S.A.*, 107(23), 10602–10607.

[207] Zhai, W., Nielsen, R., Goldman, N., & Yang, Z. (2012). Looking for Darwin in genomic sequences—validity and success of statistical methods. *Mol. Biol. Evol.*, 29(10), 2889–2893.

[208] Zhang, X.-S. & Hill, W. G. (2005). Genetic variability under mutation selection balance. *Trends in Ecology & Evolution*, 20(9), 468–470.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (XII) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.